Leveraging large scale Web data for image retrieval and user credibility estimation

Alexandru Lucian Ginsca

THÈSE / Télécom Bretagne
sous le sceau de l'Université européenne de Bretagne

pour obtenir le grade de Docteur de Télécom Bretagne
En accréditation conjointe avec l'Ecole Doctorale Sicma
Mention : Sciences et Technologies de l'Information et de la Communication

présentée par
Alexandru Lucian Ginsca
préparée dans le département Informatique

Leveraging Large Scale Web Data for Image Retrieval and User Credibility Estimation

Thèse soutenue le 30 novembre 2015
Devant le jury composé de :

Pierre-François Marteau
Professeur, Irisa/Ensibs - Université de Bretagne-Sud / président

Céline Hudelot
Maître de conférences (HDR), Centrale Supélec- Châtenay Malabry / rapporteur

Stéphane Marchand-Maillet
Professeur, Université de Genève / rapporteur

Adrian Popescu
Chercheur, CEA LIST – Palaiseau / examinateur

Ioannis Kanellos
Professeur, Télécom Bretagne / directeur de thèse

N° d'ordre : 2015telb0387

Sous le sceau de l'Université européenne de Bretagne

Télécom Bretagne

En accréditation conjointe avec l'Ecole Doctorale Sicma

Leveraging large scale Web data for image retrieval and user credibility estimation

---

Thèse de Doctorat

Mention : Sciences et Technologies de l'Information et de la Communication

Présentée par **Alexandru Lucian Ginsca**

Département : Informatique

Directeur de thèse : Ioannis Kanellos

Soutenue le 30 novembre 2015

Jury :

Mme. Céline Hudelot, MCF-HDR, École Centrale Paris (Rapporteur)
M. Stéphane Marchand-Maillet, Professeur, Université de Génève (Rapporteur)
M. Pierre François Marteau, Professeur, Université de Bretagne Sud (Président)
M. Ioannis Kanellos, Professeur, Télécom Bretagne (Directeur de thèse)
M. Adrian Popescu, Dr, CEA LIST (Encadrant)

Abstract

While research in visual and multimedia recognition and retrieval has significantly benefited from manually labeled datasets, the availability of such resources remains a serious issue. Manual annotation is still a cumbersome task, especially when it is conducted on large datasets. A promising way to circumvent the lack of annotated data is to use images shared on multimedia social networks, such as Flickr. One of the main drawbacks of user-contributed collections is that a part of images annotations is not directly related to the visual content, rendering them less useful for image mining. The work presented in this Thesis is placed at the crossroads between the use of Web data in image mining and source credibility in image sharing platforms. It aims at bringing novel findings to both domains and furnishing a promising link between two separate fields of research. The theoretical frameworks and experimental results we detail can benefit both i) researchers coming from the multimedia mining community, by introducing efficient semantic image representations built from freely available image resources and ii) researchers interested in Web data quality and source credibility, by proposing a study of credibility in the multimedia domain and testing practical applications of user credibility estimates. We propose a scalable image classification framework that exploits binary linear classifiers. To implement this framework, we compare two data sources: a large manually annotated image dataset (i.e. ImageNet) and Flickr groups. For the second, we details methods that reduce the noise inherent to a Web collection. In an extended experimental section, we show that the proposed semantic features not only improve the retrieval performance on three well known image collections (ImageCLEF Wikipedia Retrieval 2010 Collection, MIRFLICKR, NUS-WIDE), when compared to state of the art image descriptors, but also offer a significant improvement of retrieval time. We then define the concept of user tagging credibility and apply it to Flickr users. We propose 66 features that can serve as estimators for user credibility. We introduce both context and content based features extracted from various Flickr data. We evaluate the proposed features both on a publicly available dataset and new dataset, which we introduce in this Thesis. Finally, we showcase the use of credibility estimates in two application scenarios: embedding them in an image diversification pipeline and using them as features in machine learning models for expertise classification and expert retrieval tasks. This work contributes to a better understanding and modeling of social intelligence for information processing tasks. We focused on image retrieval and multimedia credibility estimation but the methods proposed here are also relevant for other applications, such as image annotation and Web data quality control.

Resumé

Au cours des années passées, l'augmentation rapide de popularité des appareils photo numériques et, notamment, smartphones, a produit de grandes quantités de données multimédias groupées dans des collections multimédias personnelles. Avec l'apparition des réseaux sociaux avec la fonctionnalité de partage de vidéos et images, tels que Flickr, Instagram, Facebook ou Youtube, ces données visuelles sont facilement partagées avec d'autres utilisateurs; une telle pratique a rapidement mené aux dépôts visuels énormes et continuellement grandissants.

L'indexation, *la recherche* et l'assurance de *qualité* de telles grandes données visuelles produites par des utilisateurs défient des problèmes qui ont besoin d'être soigneusement adressés. Ces processus exigent des méthodes d'organisation de données automatiques qui peuvent traiter efficacement de grandes quantités de données. L'approche principale au traitement de grands volumes de données multimédias compte toujours fortement sur les données textuelles (par ex. les étiquettes, les titres, la description) associées aux images. Le premier problème réside dans la qualité d'associations texte-image/vidéo. Deuxièmement, dans beaucoup de cas, les informations textuelles ne sont pas présentes ou sont rares. Dans les plates-formes telles que Flickr ou Instagram, il n'est pas obligatoire pour les utilisateurs de fournir des étiquettes à leurs contributions visuelles. Comme une alternative ou un complément à l'étiquetage manuel, beaucoup de travails se sont concentrés sur la description automatique du contenu d'image [1]. Dans cette approche, le contenu est transformé en représentations vectorielles de pixels, qui sont après utilisées pour la recherche ou la classification. La différence, du point de vue de la compréhension humaine, entre la représentation visuelle d'image et celle textuelle, est connue communément comme *le fossé sémantique* (c.-à-d. le manque de coïncidence entre les informations que l'on peut extraire à partir des données visuelles et l'interprétation que les mêmes données ont pour un utilisateur dans une situation donnée [2]). Une approche prometteuse d'adresser ce deuxième problème (c.-à-d. réduire le fossé sémantique) est d'utiliser des prédictions des détecteurs d'objets individuels ou des classifieurs en tant que descripteurs d'image [3–5]. Cette alternative devient plus complexe quand se passe à une plus grande échelle, vue que le nombre d'images devient prohibitif (c.-à-d. des centaines de millions). De plus, le nombre des concepts sémantiques qui ont besoin d'être couverts est également élevé (c.-à-d. de l'ordre des dizaines de milliers).

Dans cette thèse, nous abordons les questions susmentionnées en exploitant l'intelligence sociale dans le contexte des collections d'images à large échelle à travers du point de vue des deux domaines de l'informatique:

- **Informatique sociale:** Nous introduisons le concept de *crédibilité* dans les plates-formes de partage d'image, nous proposons un grand ensemble d'estimateurs de crédibilité pour le style étiquetage d'utilisateurs et nous montrons leur utilité pratique dans de différentes tâches liées au traitement d'image.
- **Vision par ordinateur:** Nous proposons un descripteur d'image qui transmet le sens sémantique sans utiliser des données manuellement étiquetées. Nous assurons une grande couverture de l'espace conceptuel et illustrons l'efficacité de nos descripteurs du point de vue de la grandeur, la performance de recherche et la vitesse.

Notre but est de commencer l'estimation de crédibilité à partir des morceaux de données simples et agréger ces morceaux individuels dans les estimations de la crédibilité des utilisateurs. Finalement, nous exploitons des scores de crédibilité des utilisateurs dans la recherche d'images et la classification d'utilisateurs crédibles.

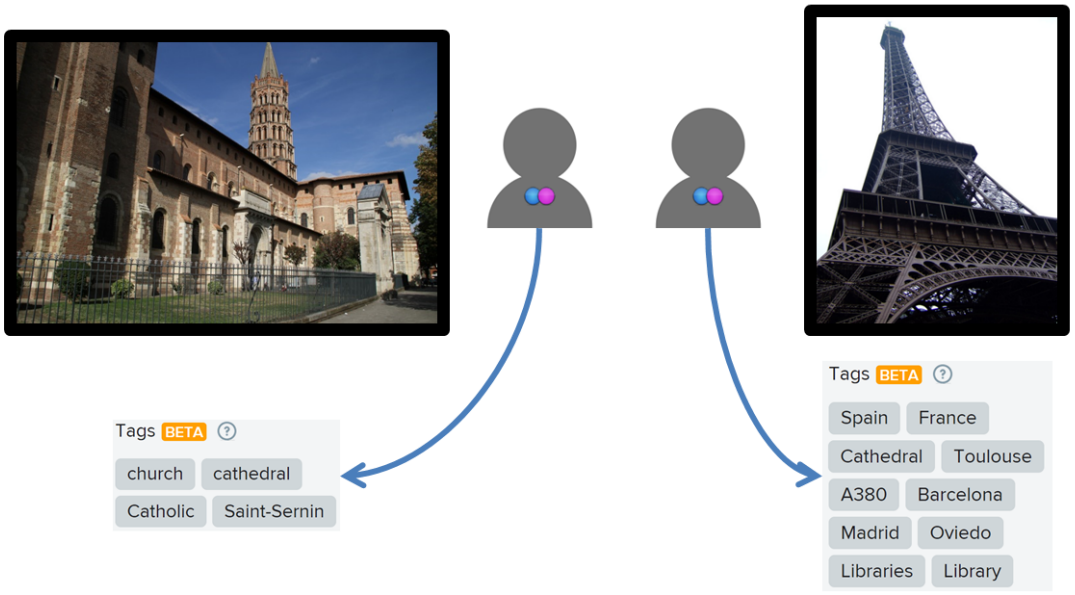


FIGURE 1: Exemples d'étiquetage sur Flickr.

Dans une des premières études sur la qualité d'étiquettes, Kennedy et al. [6] montrent que les étiquettes fournies par les utilisateurs Flickr sont extrêmement bruitées et seulement environ 50% des étiquettes sont en fait reliées aux contenu de l'image. Plus récemment, Izadinia et al. [7] ont étudié les étiquettes de 269 642 images Flickr, appartenant au jeu de données NUS-WIDE [8], pour les 81 thèmes manuellement étiquetés et ont remarqué qu'une étiquette a seulement une chance de 62% d'être correctement associée aux images. Dans une étude d'enquête, Li et al. [9] constatent que les étiquettes fournies par les utilisateurs ne peuvent pas souvent rencontrer les normes de haute qualité liées à l'association

de ce type de contenu. En particulier, ces auteurs identifient un des problèmes affectant l'étiquetage de médias sociaux, c.-à-d. le fait que les étiquettes d'utilisateur peuvent être influées par les perspectives et idées personnelles. Ainsi, les étiquettes rattachées à un contexte spécifique peuvent être préférées, ayant pour résultat souvent les étiquettes qui sont hors de propos au contenu d'image. Dans la Figure 1, nous présentons un exemple réel de deux images de Flickr et leurs étiquettes associées. Nous pouvons observer là deux types de comportement d'étiquetage. Pendant que, pour l'image gauche, les étiquettes sont pertinentes pour le contenu d'image, pour l'image à droite, la plupart des étiquettes mises par l'utilisateur sont clairement sans rapport à l'objet représenté. En estimant la crédibilité d'utilisateur, nous nous intéressons à distinguer des utilisateurs qui fournissent des étiquettes régulièrement et ceux qui utilisent les étiquettes surtout pour leur propre usage. Les contributions des derniers ne sont pas socialement pertinentes et ne devraient pas être avancées dans les applications de traitement d'image qui sont destinées pour un usage par la communauté entière.

Une deuxième contribution importante de cette thèse est liée à la description sémantique du contenu d'images. Comme prédite, il y a quelques années [10], la recherche dans la reconnaissance visuelle et multimédia a profité fortement de la disponibilité des collections d'images et de vidéos à large échelle manuellement labélisées. Avec les avances théoriques [11] et le hardware efficace, de telles ressources ont permis l'apparition de la reconnaissance visuelle, basée sur les réseaux neuronaux convolutionnels (CNN), le représentant principal d'approches de type "d'apprentissage profond". Par exemple, la représentation d'ImageNet [12] de presque 22,000 concepts, avec environ 14 millions d'images, selon une hiérarchie de concepts, a été minutieusement exploitée pour apprendre des représentations d'images puissantes et a mené à un nouvel état de l'art dans la classification d'images [13]. Pendant que puissantes, les approches "d'apprentissage profond" lèvent de nouveaux problèmes, en particulier rattachés à la disponibilité des ressources de base. Effectivement, la plupart des grandes collections nécessaires pour apprendre sont souvent manuellement labélisées, à la suite des efforts soutenus fournis par les communautés motivées de chercheurs [14] et éventuellement complétées par crowdsourcing [10], [12]. Dans ce dernier cas, une procédure de contrôle est tenue d'évaluer la qualité de l'annotation, en rendant le processus entier encore plus ennuyeux et plus long [12]. Comme principe de base, l'annotation manuelle est une tâche répétitive qui a tendance à démotiver les annotateurs ou les rendre moins précis. Finalement, crowdsourcing a un coût financier non-négligeable, quand il est conduit sur une large échelle et le financement consacré est difficile à obtenir. Une façon prometteuse de circonvenir le manque de données annotées est d'utiliser des images partagées sur les réseaux sociaux multimédias (OSNs), tels que Flickr. Un avantage de ce type de ressource comparée

aux “tâches d’annotation formelles” consiste en ce que les données sont annotées par une communauté d’utilisateurs motivés pour rendre leur contenu accessible [15].

Nous avons choisi d’utiliser la plate-forme Flickr pour l’étude d’utilisateurs crédibles, telle la source de données de Web pour la construction des concepts visuels, premièrement parce qu’elle offre un grand et divers volume de données Creative Commons (c.-à-d. images, données textuelles, les métadonnées, le réseau). Deuxièmement, un grand nombre de campagnes d’évaluation et de collections sont basées sur des données Flickr (par ex. Flickr1M [16], NUS-WIDE [17], Paris500k [18], FlickrLogos-32 [19], MediaEval Placing Task [20]). Récemment, Flickr a publié YFCC [21], la plus grande collection d’images disponibles à ce jour (99.3 millions d’images et 0.7 millions de vidéos, tous de Flickr et tous sous licence Creative Commons).

Modélisation de concept visuel à large échelle

Nous proposons un cadre de classification d’images évolutif, qui exploite des classifieurs linéaires binaires. Pour implémenter ce framework, nous comparons deux sources de données: un grand jeu de données d’images manuellement annotées (c.-à-d. ImageNet) et les groupes de Flickr. Comme la deuxième ressource est recueillie des images de Web, une partie méthodologique indispensable de travail détaille des méthodes qui réduisent le bruit inhérent à la collection. Nous fournissons aussi une évaluation préliminaire de modèles visuels individuels construits à partir de ces deux ressources. Nous enquêtons sur l’influence du nombre des exemples d’apprentissage négatifs sur la performance de prédiction et le temps d’apprentissage.

Dans cette thèse, nous explorons aussi une utilisation originale des classifieurs de concepts visuels individuels. Nous nous intéressons à enquêter le lien entre des étiquettes fournies par des utilisateurs et le contenu visuel d’une image comme une mesure sur la qualité d’étiquetage. Étant donné que dans une plate-forme de partage d’image, telle que Flickr, le vocabulaire d’étiquettes couvre un très grand nombre de concepts, les images de Flickr sont une alternative viable pour apprendre des classifieurs de concepts visuels. Nous avons donc besoin de déplacer notre intérêt des ressources manuellement étiquetées vers le fait d’exploiter le contenu produit par des utilisateurs pour construire des modèles de concepts. L’utilisation des ressources manuelles, où les concepts sont bien définis à l’avance, ainsi que les ensembles des données de Web, couplée avec une technique de réduction de bruit, mène aux éléments de base efficaces, qui peuvent servir l’apprentissage de concepts visuels. Les concepts provenus des images de Flickr sont plus proches sémantiquement aux comportements d’étiquetage réel d’ utilisateurs.

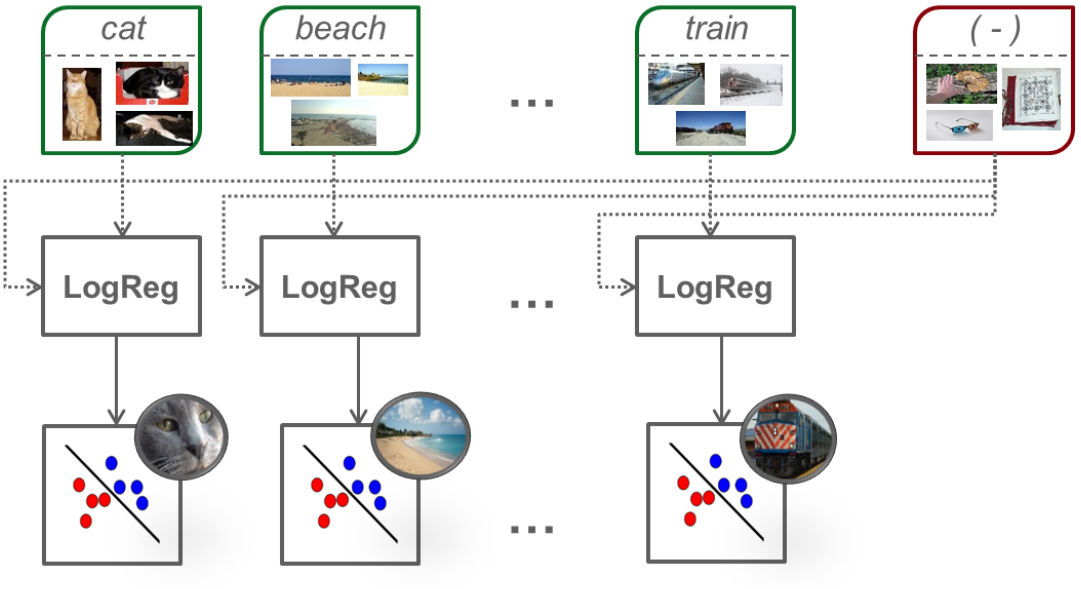


FIGURE 2: Framework d'apprentissage pour des modèles de concepts visuels.

Dans la Figure 2, nous montrons une représentation graphique du cadre proposé pour obtenir un ensemble de classifieurs de concepts visuels. Ce cadre peut être utilisé avec les données d'apprentissage positives venant de collections manuellement étiquetées ou de données de Web. L'utilisation directe de corpus de Web pour la fouille d'images, comme celle proposée dans [22] ou [23], produit une performance inférieure comparée aux jeux de données manuellement labélisées. Cependant, nous montrons qu'avec un choix approprié de la collection initiale et avec l'introduction de techniques de reclassement d'images efficaces, les résultats obtenus avec la ressource automatiquement construite peuvent égaler ceux de la ressource manuelle. Une bonne couverture de l'espace conceptuel est obtenue avec un choix approprié du jeu de données de Web. Nous avons exploré l'utilisation de groupes de Flickr, mais le pipeline présenté est facilement applicable aux collections plus grandes. Les seules contraintes potentielles sont la disponibilité de données et le pouvoir de traitement. Nous avons aussi enquêté sur le choix de descripteurs d'images et le nombre d'exemples de négatifs utilisés pour l'apprentissage des modèles.

Recherche d'images par le contenu efficace avec des descripteurs sémantiques

La contribution principale de cette partie est une approche de concevoir des descripteurs d'images sémantiques(*Semfeat*), basée sur une gamme de classifieurs de concepts individuels construits à partir des collections d'images à large échelle automatiquement traitées. Dans une section expérimentale détaillée, nous montrons que le descripteur proposé améliore la performance de recherche sur trois collections d'images bien connues (ImageCLEF 2010 Wikipedia Retrieval, MIRFLICKR, NUS-WIDE), par rapport à quelques descripteurs d'images largement utilisés. Nous montrons notamment que la

réduction de la grandeur des descripteurs sémantiques par éparsification non seulement augmente la performance de recherche, mais accélère le processus de recherche par la représentation de la collection d'images avec un index inversé.

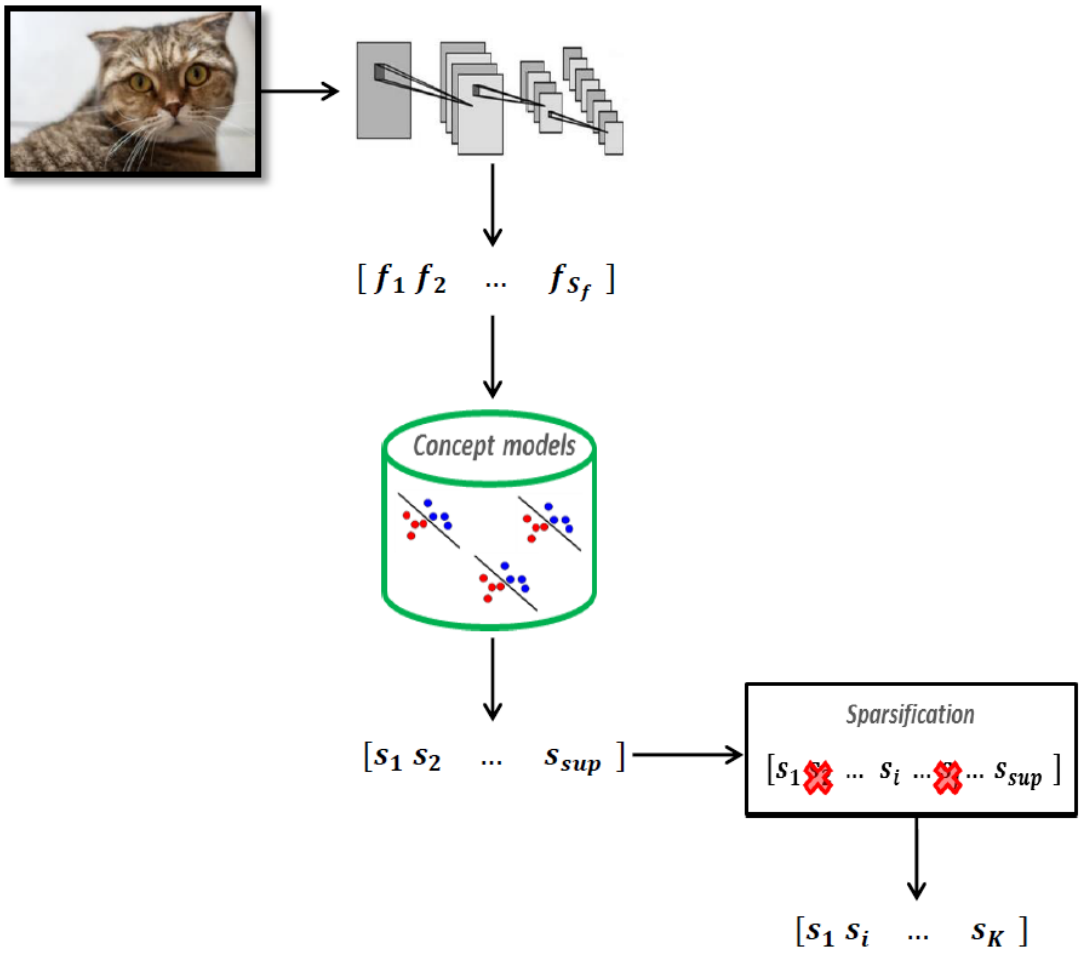


FIGURE 3: Illustration de la processus d'extraction du *Semfeat*.

Dans la Figure 3, nous illustrons l'extraction et le pipeline de sparsification pour obtenir le descripteur sémantique proposé. Pour chaque image, nous extrayons d'abord le descripteur de niveau bas. Avec l'exception unique de Fisher [24], qui est utilisé dans des expériences préliminaires, toutes représentations d'images initiales sont des descripteurs CNN (Overfeat, Caffe ou VGG). Dans l'étape suivante, nous utilisons la matrice de poids du concept (W) appris pour obtenir la représentation de *Semfeat* dense. Ce pipeline est indépendant du descripteur initial utilisé et de la collection d'images utilisée pour apprendre les concepts visuels. Nous avons évalué des configurations multiples de descripteurs sémantiques pour la recherche d'images par le contenu (CBIR). Finalement, pour obtenir une représentation compacte de Semfeat, nous gardons que les valeurs plus élevées, en mettant les restes à 0. Les résultats de CBIR montrent que les versions de Semfeat basées

sur les groupes de Flickr reclassés dépassent d'autres méthodes existantes, qui ont été essayées sur les collections de tests mentionnées.

À côté des performances compétitives et contrairement aux traits d'image largement utilisés, tels que les sacs de mots visuels, Fisher Kernels [25] ou aux descripteurs CNN [14], *Semfeat* transmet directement le sens sémantique. Les similarités d'image sont basées sur la comparaison de dimensions humainement compréhensibles (c.-à-d. les groupes de Flickr ou les concepts d'ImageNet), une caractéristique qui permet l'exploitabilité des résultats. Étant donnés une question et un résultat, les utilisateurs peuvent parcourir la liste de concepts communs pour avoir l'aspect sémantique de l'image. Ainsi, *le fossé sémantique* est réduit.

Un autre avantage de *Semfeat* est son sparsité. Les meilleures performances sont obtenues quand seulement quelques dizaines de concepts sont gardées pour chaque image. Dans cette configuration c'est facile à représenter des images à travers des index inversés pour accélérer la recherche. Nous avons évalué la recherche inversée dans la mémoire avec une mise en œuvre simple en C ++ et des collections simulées jusqu'à 1 milliard d'images. Le temps de recherche grandit linéairement et il est sous 1 milliseconde pour 10 millions d'images et sous 10 millisecondes pour 100 millions. Pour comparaison, nous avons aussi évalué la recherche directe avec Overfeat (4096 dimensions) et nous avons obtenu un temps de recouvrement dans la gamme de 15 secondes pour 10 millions d'images. Même si on utilisait des versions comprimées de traits denses, la recherche inversée serait plus rapide.

Enfin, *Semfeat* est construit au-dessus d'un ensemble de données extraites automatiquement. Nous utilisons délibérément des techniques simples, mais efficaces pour reclasser des images et apprendre les modèles. Le pipeline proposé facilite l'extension des ressources, avec la seule limitation étant la disponibilité de suffisantes ensembles d'images pour de nouveaux groupes ou concepts.

La crédibilité des utilisateurs dans les plates-formes de partage d'image

Nous définissons d'abord le concept de *crédibilité* dans les plates-formes de partage d'image. Nous enquêtons sur l'utilisation des traits de contexte et contenu pour l'estimation de la crédibilité des utilisateurs de Flickr. Nous proposons et évaluons 66 estimateurs de crédibilité. Nous faisons la fouille essentiellement sur le contenu produit par un utilisateur —principalement étiquettes et images. Nous proposons un estimateur de crédibilité visuel qui permet d'évaluer la pertinence des étiquettes associées aux images d'un utilisateur en utilisant les modèles de concepts visuels. En plus de ceux-ci, nous proposons des traits des sources de données diverses, dont un utilisateur de Flickr a des contributions — tels que les groupes de Flickr, les photos préférés, les photosets d'utilisateur ou le réseau

de contacts d'utilisateur. Pour l'ensemble des traits extraits, nous décrivons le processus d'acquisition de données et testons leur utilité en tant qu'estimateurs de crédibilité individuels. Nous définissons aussi un problème de prédiction de crédibilité, dans lequel nous apprenons des modèles de régression, qui fournissent des estimateurs de crédibilité meilleurs que les traits individuels. Nous avons constaté que, bien que les traits de contexte individuels soient de faibles indicateurs pour la crédibilité, en choisissant le modèle de régression approprié et le bon ensemble de traits pour l'apprentissage, nous sommes capables de prédire un score de crédibilité qui est considérablement meilleur corrélé avec un score de crédibilité manuel que n'importe lequel des traits individuels.

Outre investiguer sur l'utilité des estimations de crédibilité proposées sur une collection spécifique au domaine accessible librement en ligne, *Div150Cred*, nous avons également introduit un ensemble de données nouveau pour l'évaluation de la crédibilité (*MTTCred*). Après avoir décrit le processus du derrière la construction de ce jeu de données, nous fournissons des informations détaillées sur le processus d'annotation, les scores d'accord entre évaluateurs et comment nous construisons un score vérité terrain pour la crédibilité.

Nous considérons 4 composants pour la crédibilité. Selon le cas, on peut définir la crédibilité à travers d'un ou plusieurs de ces composants. Dans cette thèse, nous nous concentrons sur la crédibilité de l'utilisateur dans des plates-formes de partage d'images et, dans ce contexte, la crédibilité de l'utilisateur se reflète principalement dans la qualité des contributions d'un utilisateur. Nous identifions chaque composant de crédibilité pour la crédibilité comme il suit:

- **confiance:** comment un utilisateur est perçu par la communauté. Les indicateurs de confiance peuvent inclure le nombre d'utilisateurs qui l'ont parmi leurs contacts ou les commentaires que l'utilisateur reçoit pour ses photos.
- **expertise:** l'expertise en photographie ou la validation reçue par la communauté. Les indicateurs d'expertise peuvent inclure des indices à partir de la description de l'utilisateur (par ex. travaillant pour une institution de photographie professionnelle) ou étant invité aux groupes de Flickr exclusifs.
- **qualité:** pour la facette de crédibilité approchée dans cette thèse, nous examinons la qualité d'étiquetage d'images et pas les photos eux-mêmes. En imposant cette restriction pour le terme qualité, nous considérons qu'une image a des étiquettes de bonne qualité si elles sont bien corrélées avec le contenu visuel de l'image. Nous notons ici la différence avec la notion de véracité. Par exemple, un utilisateur peut étiqueter ses images avec le type d'appareil photo avec lequel elles ont été prises ou la date quand les photos ont été prises. Bien que pertinentes pour l'utilisateur, ces

étiquettes ne servent à aucun but pour décrire le contenu de l'image et ne peuvent pas être utilisées dans un scénario de recherche.

- **intégrité:** La qualité d'étiquetage des images d'un utilisateur (suite à la définition présentée au-dessus) est constante dans le temps.

Aux côtés de sa fonctionnalité principale de stockage de photo et de partage, Flickr fournit à ses utilisateurs des moyens d'organiser leur collection de photo, mais aussi interagir réciproquement entre eux. En plus d'étiquetage, les utilisateurs peuvent grouper leurs photos dans des photosets et peuvent ajouter leurs photos aux groupes qui reçoivent des contributions de différents utilisateurs avec un intérêt commun pour le même thème. Les utilisateurs peuvent avoir des contacts et sont capables de fournir des réactions aux photos d'autres membres de la communauté, par le biais de l'utilisation de marquages de type "préférés" et des commentaires. Nous nous intéressons à exploiter autant des données que possible pour identifier des traits d'utilisateur qui peuvent être de bons indicateurs pour la crédibilité. Les traits peuvent être groupés dans des familles de traits, selon la nature des données dont ils sont extraits. En raison des restrictions imposées par le nombre d'appels par jour aux APIs de Flickr, nous traitons les familles de traits suivantes : métadonnées de photo, groupes, photosets, des photos préférés et des contacts. Toutes les expériences et les analyses menées peuvent être facilement étendues pour inclure des traits venant d'autre source de données, quand ils deviennent disponibles. Les traits que nous extrayons peuvent être explicites et peuvent venir des actions directes d'un utilisateur, telle que l'addition d'un nouveau contact ou, implicite, que nous tirons de l'activité d'un utilisateur, tel que l'aspect temporel de son comportement de téléchargement.

Nous examinons 66 indicateurs pour la crédibilité des utilisateurs, mais nous accordons une attention particulière aux signaux visuels. Cet aspect est spécifique aux images et n'a pas été auparavant adressé dans des études de crédibilité. Dans le contexte de plates-formes de partage d'image, nous avons défini la crédibilité essentiellement par le concept de qualité. En regardant ses contributions, nous considérons qu'un utilisateur de Flickr est crédible s'il est un expert en étiquetage qui fournit des annotations de haute qualité et fiables aux photos partagés sur la plate-forme. La qualité d'une liste d'étiquette est objectivement évaluée dans les égards au contenu de l'image associée et pas au contexte dans lequel l'utilisateur a fourni les étiquettes. Dans ce sens, une liste d'étiquettes de haute qualité est celle dans laquelle les étiquettes individuelles peuvent être identifiées dans l'image et peuvent être facilement jugées comme pertinentes par d'autres utilisateurs, étant utiles ainsi pas seulement pour un seul utilisateur, mais aussi pour la communauté entière et peuvent être correctement indexées par un système de recherche. Nous fournissons une liste de traits qui peuvent être extraits directement du contenu principal des contributions d'un utilisateur sur Flickr (c.-à-d. les images et

les étiquettes). Nous exposons d'abord le processus d'extraction des traits en détail et évaluons les estimateurs de crédibilité proposés en utilisant la corrélation de Spearman avec le score de crédibilité de vérité terrain. En plus des tests sur la collection *MTTCred*, nous extrayons aussi des traits pour *DIV150Cred*.

Utilisations pratiques d'estimateurs de crédibilité des utilisateurs

Nous proposons d'abord une exploration de l'introduction des estimations de la crédibilité d'utilisateurs des systèmes de recherche d'images. Les résultats d'évaluation montrent que la crédibilité est un bon complément aux méthodes textuelles et/ou l'analyse de contenu visuelle. Les estimations de crédibilité ont été intégrées avec un algorithme de regroupement classique. Les augmentations de performance obtenues par le biais de l'utilisation de crédibilité montrent son utilité dans la recherche. Finalement, une complexité supplémentaire est ajoutée au cadre de recherche, mais affecte seulement les étapes de recherche qui sont exécutées hors ligne. Toutes les étapes peuvent être répétées périodiquement pour suivre l'évolution de la collection. Au moment de la requête, seulement un reclassement d'images qui reflète la crédibilité est exigé et cette procédure a des effets négligeables si l'on compare avec le regroupement.

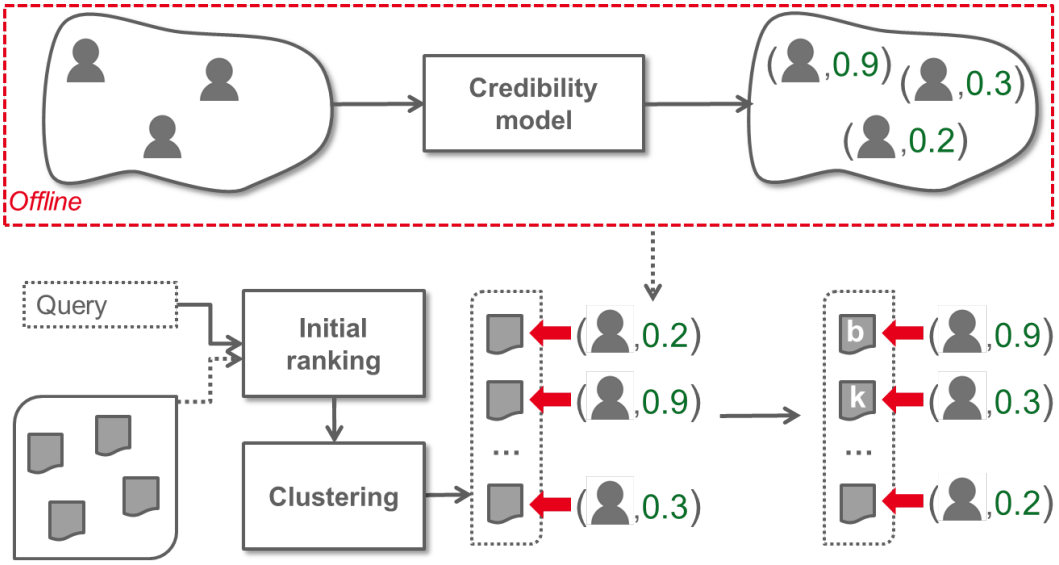


FIGURE 4: Méthode de recherche qui diversifie des images en utilisant des estimations de crédibilité des utilisateur.

Nous proposons une méthode de recherche qui diversifie des images en utilisant l'algorithme de regroupement k-Means et améliore la pertinence avec les estimations de crédibilité. Dans le cadre exposé en détail dans la Figure 4, l'estimation de crédibilité d'un utilisateur peut être tout trait de crédibilité individuel ou traits appris.

Nous proposons également deux cas d'utilisation originaux pour les estimations de la crédibilité des utilisateur. Nous enquêtons sur la pertinence des traits de crédibilité

proposés sur une tâche de classification multi-classe supervisée, adaptée à la crédibilité des utilisateur et sur une tâche de recherche, inspirée par les travaux de recherche des experts, dans lesquels nous classons des utilisateurs par leur scores de crédibilité prédits. Nous sommes aussi intéresés de mettre en évidence l'utilité de la nouvelle collection développée dans le contexte de la thèse, sur ces deux tâches. Pendant que le but principal est de faire l'évaluation sur notre collection de tests, nous utilisons le jeu de donne *Div150Cred*.

En conclusion, en cette thèse, nous montrons qu'il est possible de traiter l'intelligence sociale dans le contexte de collections d'images à large échelle, en menant à une meilleure compréhension de la crédibilité d'utilisateurs et améliorant la recherche d'images du point de vue de la qualité, la vitesse et la diversité. Les résultats prometteurs annoncés ici ouvrent un grand nombre de perspectives des travaux futurs.

Acknowledgements

I am grateful and indebted to many people who both directly and indirectly contributed to this thesis. First of all, I thank my advisers, Ioannis Kanellos and Adrian Popescu, for their continuous support during my years as their student. Ioannis's kindness, optimism and expertise pushed me to take full advantage of my PhD experience. I am also very grateful to Adrian, who stood by my side with openness, enthusiasm, and scientific knowledge. Their guidance not only had a strong impact on the quality of my work, but also on my personal development. Special thanks are in order for Adrian Iftene, who supervised my first steps in the world of research. I would also like to thank my parents, who relentlessly encouraged me to pursue my dreams. Finally, I am obliged to thank all the people I had the chance to collaborate with during these last few years: Hervé Le-Borgne, Mihai Lupu, Bogdan Ionescu, Phong D. Vo, Nicolas Ballas and Morgane Marchand.

Dedicated to the loving memory of my father.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the past years, the rapid popularity increase of digital cameras and, more notably, smartphones, produced large amounts of multimedia data in the form of personal multimedia collections. With the emergence of social networks with image and video sharing features, such as Flickr, Instagram, Facebook, or Youtube, these visual data are easily shared with other users; such a practice quickly led to enormous and continuously growing visual repositories.

Instagram[1] launched in October 2010 and is illustrative of the success of the visual-centered social networks. As of the beginning of 2015, over 20 billion photographs have been shared on the site by over 300 million monthly active users[2]. In an article published in July 2015, it has been reported that as many as 8% of Instagram accounts are fake spam bots[3]. This shows the credibility of image sources remains an issue and will be one of the central research questions of this Thesis. Flickr[4] which contained 6 billion images in August 2011 has extended its per user storage space to 1TB in May 2013[5]. Latest statistics from Flickr report that, starting from May 2015, the platform reached 112 million users[6] and stores over 10 billion images[7].

The *indexing*, *retrieval* and the insurance of *quality* of such user generated big visual data are challenging problems that need to be carefully addressed. These processes require automatic data organization methods that can efficiently process large quantities of data. The main approach to processing large volumes of multimedia data still heavily

[1] https://instagram.com/
[2] https://www.linkedin.com/pulse/2015-instagram-statistics-you-should-know-katy-elle-blake
[3] http://www.businessinsider.com/italian-security-researchers-find-8-percent-of-instagram-accounts-are-fake-2015-7?IR=T
[4] http://www.flickr.com/
[5] http://blog.flickr.net/en/2013/05/20/a-better-brighter-flickr/
[6] http://blog.flickr.net/en/2015/06/10/thank-you-flickr-community/
[7] http://blog.flickr.net/en/2015/05/07/flickr-unified-search/

relies on the textual information (*e.g.* tags, titles, description) associated to images. The first problem resides in the quality of text-image/video associations. Secondly, in many cases, textual information is not present or scarce. In platforms such as Flickr or Instagram, it is not compulsory for users to provide tags for their visual contributions. As an alternative or complement to manual tagging, a lot of work concentrated on the automatic description of image content [1]. In this approach, content is turned into a vectorial representation of pixels that is further used for retrieval or classification. The difference, in terms of human understanding, between the image representation and the textual one, is commonly known as the *semantic gap* (*i.e.* the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation [2]). A promising approach to address this second issue (*i.e.* to reduce the semantic gap) is to use predictions from individual object detectors or classifiers as image descriptors [3–5]. This alternative becomes more complex in the large scale setting, as the number of images is becoming prohibitive (*i.e.* hundreds of millions). Moreover, the number of the semantic concepts that need to be covered is equally high (*i.e.* in order of tens of thousands).

In this Thesis, we tackle the aforementioned issues by exploiting social intelligence in the context of large scale image collections through the perspective of two computer science fields:

- **Social computing:** We introduce the concept of *credibility* in image sharing platforms, propose a large set of user tagging credibility estimates and showcase their practical usefulness in different image-related tasks.
- **Computer vision:** We propose an image descriptor that convey semantic meaning without using manually labeled data. We assure a large coverage of the conceptual space and illustrate the efficiency of our descriptors in terms of size, retrieval performance and speed.

1.1 Motivation

Initial works on Web credibility include research on understanding users' mental models when assessing credibility and on the development and evaluation of interventions to help people better judge the credibility of online content. A new field, named captology [34], studies precisely how technology can be designed to persuade end-users. Credibility approaches inherit from captology perspectives, where a main goal is to understand how people evaluate credibility in order to help designers create websites that will appear

more credible. Automatic credibility estimation is a recent trend in Web content analysis and it is mostly applied to textual documents, such as tweets [35] or Web pages [36]. Also related is the automatic assessment of crowdsourced credibility, which is investigated in [37]. However, none of these works is focused on multimedia content and literature regarding multimedia credibility is limited. Xu et al. [38] aim to help users filter multimedia news by targeting credible content. They propose methods to evaluate multimedia news by comparing visual and textual descriptions respectively, as well as their multiple combinations. Yamamoto and Tanaka [39] have built ImageAlert, a system that focuses on text-image credibility. This line of prior research has shown that users consider many different pieces of information to help them evaluate the credibility of Web pages. Work on multimedia content credibility is at best incipient and, to the best of our knowledge, the contributions introduced in this Thesis are one of the first attempts to automatic credibility prediction for visual social network users.

A majority of researchers identify two components of credibility, namely trustworthiness and expertise [40]. In general, trustworthiness is understood as unbiased, truthful, well intentioned, while expertise is taken to mean knowledgeable, experienced, or competent. We advocate that quality and reliability are two supplementary essential aspects of the concept. Quality is seen as an intrinsic characteristic of content shared by Web users, while reliability refers to the extent to which something can be regarded as dependable and consistent. Our purpose is to start credibility estimation from single data pieces, aggregate these individual pieces into estimations of user credibility. Finally, we exploit user credibility scores in image retrieval and credible user classification and ranking tasks to showcase their practical usability.

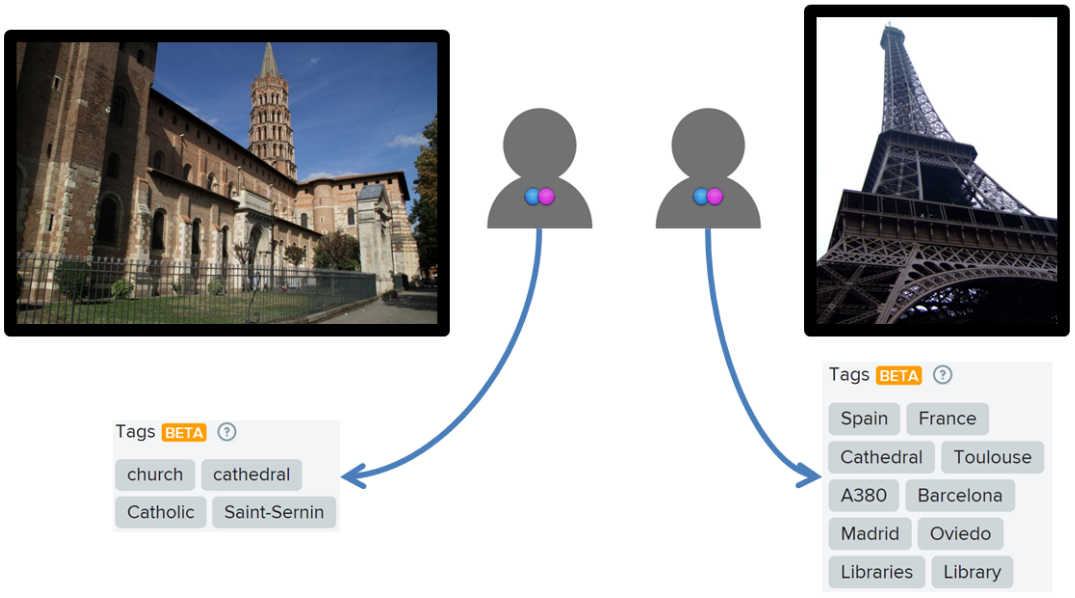


FIGURE 1.1: User tagging examples on Flickr.

In one of the first studies on the quality of Flick tags Kennedy et al. [6] show that the tags provided by Flickr users are highly noisy and there are only around 50% tags actually related to the image. More recently, Izadinia et al. [7] studied the tags of 269 642 Flickr images from the NUS-WIDE dataset [8] for the 81 manually labeled topics and observed that a tag has only a 62% chance of being correctly associated to images. In a survey paper, Li et al. [9] find that tags provided by users often cannot meet the high quality standards related to content association. In particular, these authors identify one of the problems affecting social media tagging, *i.e.* the fact that user tags may be biased towards personal perspectives and ideas. Thus, tags related to a specific context may be preferred, often resulting in tags that are irrelevant to the image content.In Figure 1.1, we present an actual example of two Flickr images and their associated tags. We can observe there two types of tagging behaviors. While for the left image, the tags are relevant for the image content, for the right image, most of the tags are clearly unrelated to the depicted object. By estimating user tagging credibility, we are interested in distinguishing between users who provide relevant tags on a regular basis and those who use tagging mostly for their own usage. The contributions of the latter are not socially relevant and should not be put forward in image-related applications that are intended for a community usage.

A second important contribution of this Thesis relates to the semantic description of image content. As predicted a few years ago [10], research in visual and multimedia recognition has strongly benefited from the availability of manually labeled large-scale image and video collections. In conjunction with theoretical advances [11] and relatively cheap and efficient hardware, such resources enabled the emergence of visual recognition based on convolutional neural networks (CNN), the mainstream representative of "deep learning" approaches. For instance the ImageNet representation [12] of nearly 22,000 concepts, with approximately 14 million images, according to a hierarchy of concepts was thoroughly exploited to learn powerful image representations and led to a new state of the art in image classification [13]. While powerful, "deep learning" approaches raise new problems, in particular related to the availability of the underlying resources. Indeed, the large datasets needed for learning most are often manually labeled, as a result of sustained efforts provided by motivated communities of researchers [14], and eventually supplemented with crowdsourcing [10], [12]. In this last case, a control procedure is required to assess the quality of the annotation, making the whole process even more tedious and longer [12]. As a rule of thumb, manual annotation is a repetitive task that tends to demotivate the annotators or make them less accurate. Finally, crowdsourcing has a non-negligible financial cost when it is conducted on a large scale dataset, and dedicated funding is difficult to obtain. A promising way to circumvent the lack of annotated data is to use images shared on multimedia social networks (OSNs), such as

Flickr. An advantage of this type of resource compared to formal "annotation tasks" is that data are annotated by a community of users motivated to make their content accessible [15].

We chose to use the Flickr platform both for the study of user tagging credibility and as Web data source for visual concept building, firstly because it offers a large and diverse Creative Commons volume of data (*i.e.* image, textual, metadata, network). Secondly, a large number of evaluation campaigns and datasets are based on Flickr data (*e.g.* Flickr1M [16], NUS-WIDE [17], Paris500k [18], FlickrLogos-32 [19], Placing Task at MediaEval [41], Retrieving diverse social images task at MediaEval [20]). Recently, Flickr publicly released YFCC [21], the largest image collection available to date (99.3 million images and 0.7 million videos, all from Flickr and all under Creative Commons licensing).

1.2 Contributions and Outline of This Thesis

The work presented in this Thesis is placed at the crossroad between the use of Web data in content based image retrieval (CBIR) and source(user) credibility estimation in image sharing platforms. It aims at bringing novel contributions to both domains and at proposing a promising link between two separate fields of research. The theoretical frameworks and experimental results that we detail can benefit both to: i) researchers coming from the multimedia retrieval community, by introducing efficient semantic image representations built from freely available image resources and ii) researchers interested in Web data quality and source credibility, by proposing a study of credibility in the multimedia domain and testing practical applications of user credibility estimates.

In Figure, 1.2, we offer reading paths to readers interested in the different aspects of our contributions. Someone coming from a computer vision background may be mainly interested in following the *several blue paths*. Our focus there is on large scale visual concept learning (Chapter 3) and semantic image representation for content based image retrieval (Chapter 4). On the other hand, if the reader is more interested about Web data quality and user expertise, he or she may want to follow the *green paths*. In the second part of Chapter 2, we define our interpretation of Web credibility and provide a thorough survey of recent works in this field. In Chapter 5, we instantiate our credibility model in the multimedia domain, building notably on 3 contributions, and in Chapter 6 we put to test the observations made in the previous Chapter in two scenarios. In this line of work, Annex 1 stands out as a particular study on the relation between user expertise and data quality in the Social Question&Answering domain. The main contributions introduced in each Chapter are summarized as hereafter.

Chapter 2. We present in this Chapter seminal works and recent advancements in both computer vision and Web credibility applications. Considering the lack of a common view on the concept of Web credibility, the larger part of the second chapter is covered by a thorough survey of credibility-related works, focusing on this notion in social networks and multimedia.

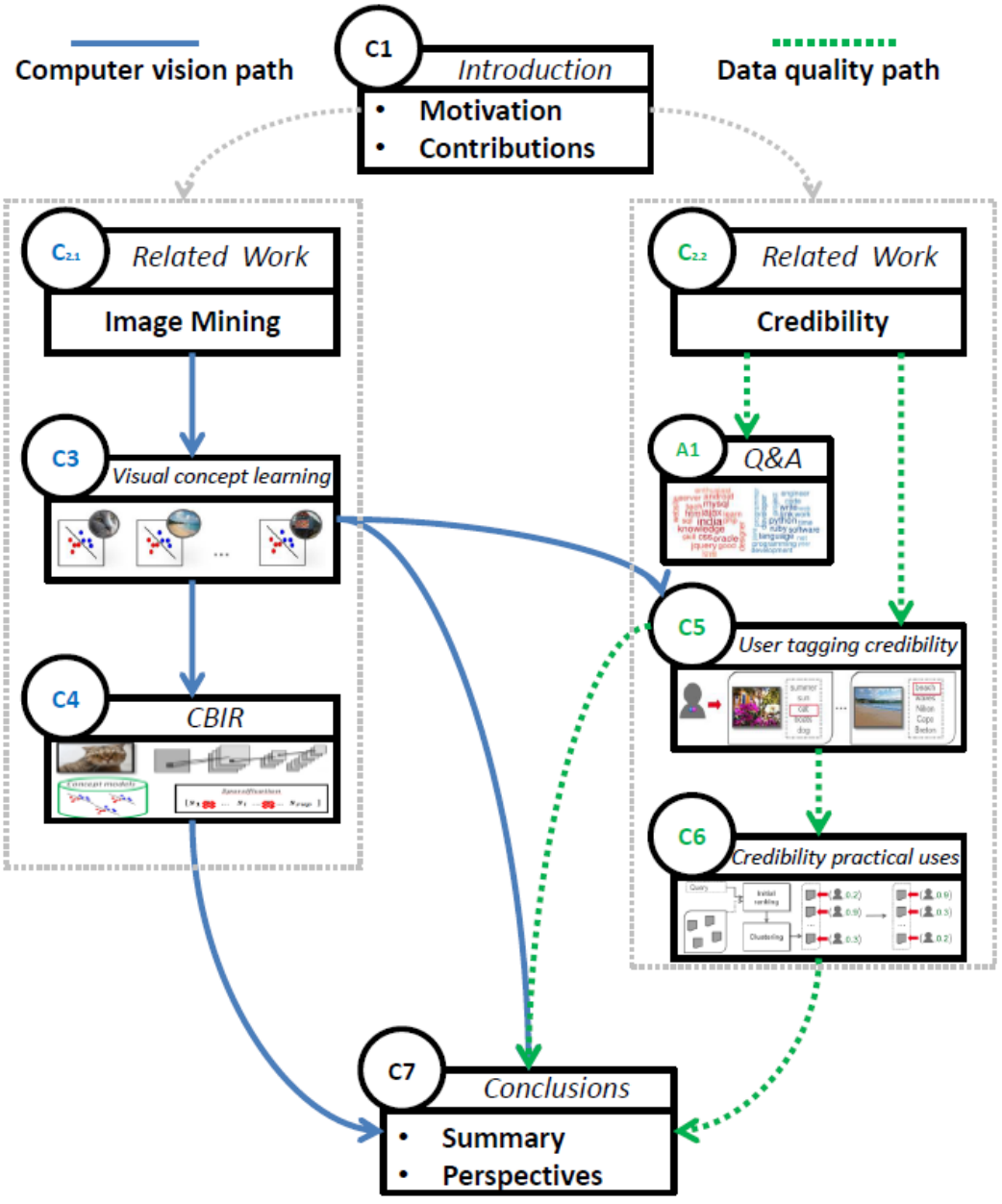


FIGURE 1.2: Thesis reading map.

Chapter 3. We propose in this Chapter the use of binary linear classifiers to train models for a large set of concepts. We compare two data sources: ImageNet, a large manually annotated image dataset, and Flickr groups images. Since the second resource is collected from Web images, a key methodological part of the work deals with some

methods that allow reducing the noise that is inherent to such a Web collection. We present preliminary results of individual classification models that support the validity of our approach. These results are then extensively exploited in the following chapters.

Chapter 4. The main contribution of this Chapter is an approach to design semantic image features that are based on an array of individual concept classifiers built on top of automatically processed large-scale image collections introduced in Chapter 3. In an extended experimental section, we show that the proposed features not only improve the retrieval performance on three well known image collections (ImageCLEF Wikipedia Retrieval 2010, MIRFLICKR, NUS-WIDE), when compared to widely used image descriptors, but also offer a significant improvement of retrieval time.

Chapter 5. Here, we first define the concept of user tagging credibility in the context of the Flickr platform. Our main goal is to propose multiple features that can serve as estimators for user credibility. We introduce both context and content features stemming from various data sources (Flickr groups, photo favorites or a user's contacts network) and also exploit the set of concept classifiers introduced in Chapter 3. Another contribution of this part is the creation of a new dataset specifically built i) to help us evaluate potential indicators for credibility and ii) to serve as a training dataset on which one can compare different learning models and features. We then detail the motivation behind the need for such a dataset and the methodology used for its creation. We also detail dataset statistics, the data acquisition process for a large set of features and then test their usefulness as individual credibility estimators. Finally, we define a credibility prediction setting, in which we learn regression models that provide better credibility estimators than the individual features.

Chapter 6. In this Chapter, we investigate the use of credibility estimates in two different scenarios. Firstly, user credibility estimates are introduced in an image search diversification task to rerank a list of retrieved items. Our approach is validated on a publicly available dataset. We then showcase the use of both the dataset and the credibility features introduced in the previous chapter in two scenarios, i.e. user credibility classification and credible user retrieval. We find that using off-the-shelf models allows the exploitation of the proposed features to accurately differentiate between credible and non-credible users and to provide a relevant user ranking.

Annex 1. This Annex is a complementary user credibility study carried on a Question&Answering platform. Our goal is to determine if there are any particular dimensions of a user's profile or activity in the community that can be exploited to spot high quality answers. We first perform an in-depth analysis of the information provided by the users in their profiles in order to discriminate features that are correlated to expertise. Then we investigate the importance of topics associated to answers based on the

community feedback a user question receives. To do this, we propose a topic model-based approach that is testes by determining the quality of newly submitted answers. Finally, we propose an answer ranking scenario in which we assess the predictive capabilities of profile and activity features and the usefulness of the best topic models methods. In our experiments, we use a large scale corpus from Stackoverflow, a very active Q&A community focused on technical topics. For answer quality prediction, we compare Latent Dirichlet Allocation and Explicit Semantic Analysis based methods to a baseline that attributes an average score to newly arrived items and show that improved predictions are obtained with both types of methods. We show that automatic answer rankings obtained by exploiting different user features outperform a natural ranking based on temporal order.

In summary, in this Thesis we show that it is possible to treat social intelligence in the context of large scale image collections, leading to a better understanding of user tagging credibility and improving image retrieval in terms of quality, speed and diversity. The promising results reported here open a number of future work perspectives that are described in the last chapter of the Thesis.

Chapter 2

Related work

In this chapter, we present seminal works and recent advancements in both computer vision and Web credibility. Considering the lack of a common view on the concept of Web credibility, the larger part of the second chapter is covered by a thorough survey of credibility related work, focusing on this notion in social networks and multimedia.

2.1 Image mining

In this Section, we first review significant and recent work in the literature on image descriptors. Then, we present advancements both in object detection and classification and content based image retrieval, with a focus on image collections representation methods.

2.1.1 Image representation

The choice of image descriptors is one of the main factors that impacts the accuracy of an image classification system [42]. For a long period of time, several techniques have been developed for representing the content of images. These can be distributed into two categories: (i) *global descriptors*: methods which directly extract pixel based features from the images, and (ii) *local descriptors and aggregates*: methods which model the image using local patches as an intermediate representation. However, with recent advancements in Convolutional Neural Networks (CNN) based image descriptors and an interest in having an image descriptor capable of directly conveying semantic meaning, we propose the following classification of image descriptors:

- low-level image descriptors: We include in this category both classical local and global image descriptors. Throughout this Thesis, we will refer to any image representation from this category as a low-level image descriptor.
- mid-level image descriptors: We refer as mid-level image descriptors Convolutional Neural Networks features extracted from the weights of intermediate layers.
- high-level image descriptors: We consider high-level image descriptors an image representation in which each component of the vector has a semantic significance. They are usually obtained through the use of object detectors of visual concept classifiers. We will also use the term "semantic features" when describing image descriptors coming from this class.

We will next cover these categories, with an emphasis on the last two, which are closer to the image representations we use or propose in this Thesis.

2.1.1.1 Low-level image descriptors

Early works described images using global signatures based on various aggregations of pixel-level statistics. The global gray-scale image histogram is an example of such an image representation. It counts the number of times a certain pixel value appears in an image. Another example of global image descriptors is the GIST descriptor [43]. GIST represents a low-dimensional description of the whole image through a set of perceptual dimensions such as naturalness, openness, roughness, expansion and ruggedness.

Local features are extracted from patches or interest points. The feature detection step determines the number, the size and the location of the patches that are extracted in an image. We mention here three main methods used for feature detection in the literature: sparse detection based on the interest points [44], detection on a dense grid [45] and random sampling of the patches [46].

We briefly present some of the most popular local features.

The SIFT descriptor [47] builds a histogram of image gradients within each patch. It computes 8 orientation directions over a 44 grid. Through a Gaussian window function that gives more weight to the gradients computed near the center of the patch, the SIFT descriptor is considered to offer robustness to some level of geometric distortion and noise. Also, for robustness to illumination changes, the SIFT descriptor is normalized to one. Color SIFT [48] computes SIFT descriptors separately for the red, green and blue channels. Speeded up Robust Features (SURF) [49] is another scale and rotation

invariant local feature extraction algorithm that computes gradients in only two orientations and relies on image integral masks to approximate the gradient computation. Local Self Similarity (LSS) [50] describes an interest point by computing the sum of squared distances between a small patch whose center is the sampled point and other patches from a bigger region.

Detection of local image regions is only the first part of the feature extraction process. The second part is the computation of descriptors to characterize the appearance of these regions. During the last decade, the most widely used image retrieval features relied on the aggregation of local features, such as bags of visual words (BoVW) [51, 52]. These approaches first extract descriptors such as SIFT or SURF from image patches and then aggregate them into a fixed size vector BoVW that describes the global properties of the image. Before the rise of CNNs, BoVW has been the major feature descriptor in many computer vision applications [53–55]. As an improvement over the BOV representation, Van Gemert et al. [54] suggest to soft-assign the local descriptors using a generative model built on the descriptors.

Several works have proposed to perform an explicit embedding of the image representations in a high dimensional space where the BoVW histograms are more linearly separable. Maji and Berg [56] proposed mappings for the Intersection Kernel (IK) and Wang and al. [57] proposed efficient algorithms to learn IK SVMs. BoVW were improved through the introduction of higher-order image statistics in features such as *Fisher vectors* (FV) [58]. While the BoVW descriptor is composed on only the count of visual word occurrences for each local descriptor, FV consists in computing the deviation of a set of local descriptors from an average Gaussian Mixture Model. Normalizing the FV improves its descriptive power [59].

A problem common to these descriptors is their high dimensionality. Different compression methods were proposed to improve scalability. *VLAD* (vector of locally aggregated descriptors) [60] successfully reduced the size of Fisher vectors and was further optimized by the introduction of *PQ* (product quantization) method [61]. With *VLAD+PQ* representation, 100 million image features would fit into 2 GB of RAM and could be searched in approximately 240 ms on a single core. While improving scalability, the aggressive compression performed by *VLAD+PQ* significantly decreases accuracy compared to the use of full FV.

2.1.1.2 Mid-level image descriptors

In the last years, convolutional neural networks (CNN) have become standard practice in many computer vision tasks. The initial breakthroughs have been lead by improved accuracies on the large-scale visual recognition challenge ILSVRC [62] with CNNs trained on the ImageNet objects categories [63]. Compared to traditional low-level features such as Fisher Vector [64], the use of CNN brought down the ILSVRC error rate from 0.26 to 0.15 in 2012, 0.11 in 2013 [14, 63] and 0.07 in 2014 [65]. Notable efforts were devoted to studies of effects of different modes of training and experimenting with different architectures [13, 66–68].

Pepik et al. [69] compiled a comprehesive list of classical computer vision problems that have now all top performing results based on a direct usage of CNNs: image classification [63], object detection [70], pose estimation [71], face recognition [72], object tracking [73], keypoint matching [74], stereo matching [75], optical flow [76], boundary estimation [77], and semantic labeling [78].

Since the initial success, CNN features have been used as universal representation for a variety of classification tasks ([79] and [80]). In addition to object categorization, the use of CNN architectures for object localization [31], scene classification and other visual recognition tasks have been demonstrated. More important, CNN-based feature extractors were publicly released. This meant that the use of CNN-based features became available without requiring the knowledge or computing infrastructure for training a convolutional neural network from scratch. Among the first tools that were made publicly available we can cite *Overfeat* [14], followed by *Caffe* [81]. These extractors provide pre-trained weights files and facilitate the extraction of features for new image collections. The outputs of their final layer are semantic image representations but they are limited to the 1,000 ILSVRC concepts, due to computational complexity of the algorithm. CNN features extracted from intermediate layers are most often referred as mid-level or intermediate image descriptors (for a visual example of CNN feature extraction, see Figure 3.1). While they do not have a semantic interpretation (as high-level descriptors), unlike low-level descriptors, they still capture information about the represented image. Depending on the position of the layer in the network architecture [67], these descriptors can give clues about texture or shapes, when using an appropriate visualization [66, 82, 83].

Several works investigated the performance of CNN features with the goal of getting better understanding of their usefulness for various classification tasks. Rigorous evaluation of the comparison of CNN methods with shallow representations such as Bag-of-Visual-Words and Improved Fisher vectors has been conducted in [84]. The evaluation was carried out on the different categorization tasks (ImageNet, Caltech and PASCAL-VOC).

The premise of this study was to compare different representations which are suitable for the analysis with linear classifiers, such as SVM. The experiments concluded that, while the shallow methods can be improved using data augmentation, the CNN representations significantly improve the classification performance. Gong et al. [85] proposed computation of CNN features over windows at multiple scales and aggregating these representations in a manner similar to Spatial Pyramid Pooling, affecting favorably both the classification and image based retrieval performance. While the pooling strategy was found effective, the features extraction stage was expensive, yielding high feature dimensionality. All the methods mentioned above used the last fully connected layer (*fc7*) features as image or window representations with dimensionality of 4096. The convolutional level 5 features have been evaluated in the absence of pooling strategies on Caltech-101 dataset in [84], yielding inferior performance compared to fully connected layer features *fc6* and *fc7*. With the exception of [85], the above mentioned studies focus on classification instead of image retrieval tasks.

Another line of work that is even to more interest for us is related to the direct use of CNN features image retrieval. Representations used in the past for the CBIR used both local and global features. They often considered as baseline method the BoVW representation, followed by spatial verification of top retrieved images using geometric constraints [86]. Various improvements of these methods include learning better vocabularies, developing better quantization and spatial verification methods [87] or improving the scalability. Alternative more powerful quantization and representation techniques have been also explored in [88–90]. Chatfield et al. [91] investigate the gains in precision and speed, that can be obtained by using Convolutional Networks (ConvNets) for on-the-fly retrieval, where classifiers are learned at run time for a textual query from downloaded images and used to rank large image or video datasets. They show that the CNN descriptors can be efficiently compressed and used in a incremental learning architecture. They conclude that the proposed architecture is capable of retrieval across datasets of over one million images within seconds and running entirely on a single GPU.

While accurate, CNN features usually have a size in the range of thousands of dimensions that makes their direct use for large-scale retrieval difficult. The authors of [92] compared *CNN*, *VLAD* and *VLAD+PQ* in an ah-hoc retrieval task on the YFCC100M collection that includes nearly 100 million images [93]. Results show that the precision of *CNN* features is roughly three times higher than that of *VLAD* and *VLAD+PQ*. Equally important, an aggressive PCA compression of *CNN* to only 16 dimensions only degrades performance to approximately $\frac{2}{3}$ of full features. Compressed forms of existing features, such as *PCA-CNN* or *VLAD+PQ*, enable real-time retrieval on a single core for collections up to 100 million images but they reduce accuracy and also require distribution on several machines for larger datasets.

TABLE 2.1: A selection of the most prominent semantic image descriptors currently available. The table presents the paper where the descriptor was introduced, different configurations and the size of the descriptor.

Reference	**Descriptor configuration**	**Descriptor size**
Li et al. [94]	KMS	14.3 K
	VQ	10 K
Bergamo and Torresani [23]	MC-LSH	200 K (binary)
	MC-BIT	15.2K (binary)
	MC	15.2K
Su and Jurie [95]	Attribute classiffiers	110
Bergamo et al. [4]	PICODES	2048 (binary)
Lin et al. [96]	ObjectBank +ASGD	1.1 M
Gong and Lazebnik [97]	CCA-ITQ	2048 (binary)
	ITQ	200 K (bin)
Torresani and al. [22]	Classemes-bit	2659 (binary)
	Classemes	2659
jia Li et al. [3]	ObjectBank	44.6 K

2.1.1.3 High-level image descriptors

The image itself, as humans perceive it, has all the essential information about the content of the image. However, as computers "perceive" it, the image itself contains only low-level information about its individual pixels. High level image descriptors can be extracted from an image to bridge the gap between low-level information and high level concepts.

The approach introduced in [22] can serve as a general framework of classifier-based image descriptors. At a high-level, extracting semantic features involves representing an image x as a k-dimensional vector $s(x)$, where the $i-th$ entry is the output of a classier or object detector C_i evaluated on x:

$$s(x) = \begin{bmatrix} C_1(x) \\ \vdots \\ C_k(x) \end{bmatrix} \tag{2.1}$$

The classifiers $C_{1...k}$ (the basis classifiers) are learned during an offline stage from a manually or automatically labeled large collection of images.

The availability of large image collections and of scalable machine learning techniques has led to a resurgence of semantic representation for image classification [3, 22, 95]. Li et al. [3] introduced Object Bank, where an image is represented as a scale-invariant response map of 200 pre-trained object detectors. In a follow-up work, in order to facilitate the training of larger set of concepts, [96] develop a Hadoop scheme that performs feature

extraction in parallel using hundreds of mappers. This allows the extraction of highly dimensional features features (hundreds of thousands) on 1.2 million images within one day. For SVM training, they develop a parallel averaging stochastic gradient descent (ASGD) algorithm for training one-against-all 1000-class SVM classifiers. An extension of object bank, called action bank [98], is proposed to represent complex activities in videos. Torresani et al. [22] also introduced a semantic representation using a fix number of hand selected binary classifiers. Each classifier is applied on the whole image input. In a closely related work, Su and Jurie [95] used 110 manually selected attributes to represent images. Due to the relatively small number of visual concepts considered, early semantic representations ensured only a limited coverage of the semantic space. To tackle this issue, Bergamo and Torresani [23] learned the visual concepts of the semantic representation directly from the data. They however use 13 different features and "lift-up" each one to approximate a non-linear kernel, that is a much more costly approach than ours. Closer relater to our use of Web images from the Flickr platform, Li et al. [94] propose a fully automatic algorithm which harvests visual concepts from a large number of Web images (more than a quarter of a million) using text-based queries. Unlike us, they use Google and Bing image data and collect images for around 14,000 visual concepts. With their best configuration, they obtain a 62.9% mean average precision on PASCAL VOC 2007. While this is an improvement over the Fisher Kernel they used for comparison [99] (59.6%) this result lags behind current CNN approaches (*e.g.* 82.4% mAP by [13]). More than that, even thou in this Thesis we focus on CBIR, in a preliminary experiment on PASCAL VOC 2007, we obtain a 73.6% mAP with our proposed semantic descriptor (see Section 4.9.2 of Chapter 4).

Bergamo et al. [4] introduced PiCoDes, a feature in which they use basis classifiers as features with linear models. They learn abstract categories aimed at optimizing linear classification when they are used as features. This learning objective decouples the number of training classes from the target dimensionality of the binary descriptor and thus it allows the optimization of the descriptor for any arbitrary length. The learned features describe the image in terms of binary visually-interpretable properties corresponding, e.g., to particular shape, texture or color patterns. The Meta-class (MC) [23] descriptor is obtained through a hierarchically partition over the set of training object classes such that each meta-class subset can be easily recognized from the others. This criterion forces the classifiers trained on the meta-classes to be repeatable. In a recent paper, Bergamo and Torresani [5] provided an overview of semantic descriptors and also proposed methods to aggregate the locally-dependent outputs of the basis classifiers into a single feature vector, thus rendering the descriptor more robust to changes in size and position of the object of interest.

More closely related to our work is the idea presented in [57], where the similarity between two images is computed by leveraging the prediction scores of a set of 103 hand picked Flickr groups. Each probability is estimated using a SVM classifier trained over low-level visual features. The resulting vectors are also briefly tested in clustering and classification tasks, for which comparable results with visual features are reported. Key differences with our work arise from the way groups are modeled. We propose several image ranking methods that improve individual classifier performance when using an initial training set of only 300 images, whereas in [57] the learning is performed on a large training set (15,000 to 30,000 images). In addition, we sparsify the features and thus enable fast retrieval over large datasets.

2.1.2 Visual concept classification

The problem of concept recognition in large datasets has been the subject of much recent work. While nonlinear classiers are often seen as state-of-the-art in terms of categorization accuracy [65, 100], they are difficult to scale to large training sets. In consequence, more efficient linear models are usually used in recognition settings involving a large number of object classes, with many image examples per class [101]. As a result, much work in the last few years has focused on methods to retain high recognition accuracy even with linear classifiers. In this Thesis, we use a large set of binary classifiers as an intermediate step for building semantic features. For this reason, we do not detail in this Subsection the highly expanding ecosystems of works regarding the use of CNNs for multi-class classification tasks.

One category of classifiers comprises techniques to approximate nonlinear kernel distances via explicit feature maps [56, 102]. For many popular kernels in computer vision, these methods provide mappings to higher-dimensional feature spaces where inner products approximate the kernel distance. However, these methods are typically applied to hand-crafted features that are already high dimensional and they map them to spaces of further increased dimensionality. A second line of work [103] involves the use of vectors containing a very large number of features (up to several millions) to obtain a high degree of linear separability. The idea is similar to that of explicit feature maps, with the difference that these high-dimensional signatures are not produced with the goal of approximating kernel distances between lower-dimensional features but rather to yield higher accuracy with linear models. These vectors are typically stored in compressed form and then decompressed quickly and one at a time during training and testing [81, 103]. As an alternative, in Lin et al. [96], the high storage costs caused by their high-dimensional descriptor were resolved by a large system infrastructure consisting of Apache Hadoop to distribute computation and storage over many machines.

Another line of related work involves the use of image descriptors encoding categorical information as features. The image is represented in terms of its closeness to a set of basis object classes [22, 57, 104] or as the response map to a set of detectors [3]. These works can be seen as a mean of using high level feature extraction inspired frameworks for classification or even directly using semantic features for this task. Unexpectedly, even linear models applied to these high-level representations have been shown to produce good categorization accuracy. These descriptors can be viewed as generalizing attributes [105–107], which are semantic characteristics selected by humans as associated with the classes to recognize.

Closely related to concept classification is the line of work involving the use of attributes [105–107] which are fully-supervised classifiers trained to recognize certain properties in the image such as *has wheels* or *has fur*. Attributes have been used as features for recognition in specialized domains (e.g., animal recognition [107] or face identification [106]). Farhadi et al. [105] worked on describing objects by parts, such as *has head*, or appearance adjectives, such as *spotty*. They distinguished between two types of attributes: semantic (*spotty*) and discriminative (*e.g.* one animal has the attribute but another don't). Similarly, [106] considered two types of attributes for face recognition: those trained to recognize specific aspects of visual appearance, such as gender or race, and simile classifiers which represent the similarity of faces to celebrity faces. Ferrari and Zisserman [108] proposed learning attributes using segments as the basic building blocks. They distinguish between unary attributes (colors) involving just a single segment and binary attributes (stripes, dots and checkerboards) involving a pattern of alternating segments. Yanai and Barnard [109] learned the *visualness* of 150 concepts by performing probabilistic region selection for images labeled as positive and negative examples of a concept, and computing the entropy measure which represents how visual this concept is. These can be seen as discriminative attributes. They evaluated their algorithm on Google search images, and also considered each image to be a collection of regions obtained from segmentation, but didn't consider the pairwise relationship between the regions. Lampert et al. [107] considered the problem of object classification when the test set consists entirely of previously unseen object categories, and the transfer of information from the training to the test phase occurs entirely through attribute text labels. They introduced the Animal with Attributes dataset with 30,000 images annotated with 50 classes. They are interested in performing zero-shot object classification based on attribute transfer rather than learning the attributes themselves or building an attribute hierarchy. Rohrbach et al. [110] use semantic relationships mined from language to achieve unsupervised knowledge transfer. They found that path length in WordNet is a poor indicator of attribute association. They show that web search for part-whole relationships is a better way of mining attribute annotations for object categories. Farhadi

et al. [111] discussed creating the right level of abstraction for knowledge transfer. They learned part and category detectors of objects, and described objects by spacial arrangement of their attributes and the interaction between them. They focused on finding animal and vehicle categories not seen during training, and inferring attributes such as function and pose. They learn both the parts that are visible and not visible in each image. We are more interested in semantic attributes, as these can serve as indicators for our line of work in building semantic features, but for a broader understanding of the field, we have also presented discriminative and comparative attributes.

Close to concept classification is a body of works that deal with object detection. The problem of object detection has been traditionally approached as the task of exhaustive sub-image recognition [112, 113]. For every category of interest, a classifier is evaluated at every possible rectangular subwindow of the image, thus performing a brute-force sliding window search. In order to maintain the computation manageable despite the large number of subwindows to consider, these approaches are constrained to use features that are extremely fast to extract (*e.g.* HOG [114] and Haar [115]). Uijlings et al. [116] and Alexe et al. [117] have identified inside the image the rectangular subwindows that are most likely to contain objects, regardless of their class. Particularly the method of selective search originally proposed in [48] shows a recall (fraction of the true objects that are identified by the method) approaching 97% for a small number of candidate subwindows (on average about 1500 per image). Also, this enables the practical application of sophisticated features and object detection models, which instead would be prohibitive in a traditional sliding-window scenario. For example, the solution proposed in [48] achieves competitive results by training a nonlinear SVM on a spatial pyramid of histograms computed from 3 distinct local appearance descriptors. Despite the complexity of this model, the computational cost of recognition remains low if the classier is applied only to the 1500 candidate sub-images rather being exhaustively evaluated over all possible subwindows. Most weakly-supervised object detection methods [118–121] aim at jointly learning and inferring both the class and the position of the objects.

In the context of both object localization and detection, more recently, researchers have applied deep networks [70, 122–128]. In [70], a convolutional network is fine-tuned on ground truth bounding boxes and then applied to classify subwindows generated by the region proposal algorithm of Uijlings et al. [116]. In [122–124] a convolutional network is trained to directly perform regression on the vector-space of all bounding boxes of an image in order to avoid the high computational cost of traditional sliding window or region proposal approaches. Ren et al. [126] introduce a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. An RPN is a fully-convolutional network that simultaneously predicts object bounds and object scores at each position. RPNs are

trained end-to-end to generate high quality region proposals, which are used by Fast R-CNN for detection. They show that with a simple alternating optimization, RPN and Fast R-CNN can be trained to share convolutional features. Similarly, Gidaris and Komodakis [127] propose an object detection system that relies on a multi-region deep convolutional neural network (CNN) that also encodes semantic segmentation-aware features. The resulting CNN-based representation aims at capturing a diverse set of discriminative appearance factors and exhibits localization sensitivity that is essential for accurate object localization. Both of the previous presented methods obtain high scores on the detection challenges of PASCAL VOC2007 and PASCAL VOC2012 [129] (Ren et al. [126] report a mAP score of 73.2% on VOC2007 and 70.4% on VOC2012, while Gidaris and Komodakis [127] achieve mAP of 78,2% on VOC2007 and 73,9% on VOC2012). These deep networks have shown promising results compared to standard detection schemes relying on handcrafted features [112, 116]. However, nearly of all them require manually-annotated ground truth bounding boxes as training data.

Although we do not use object detectors for the semantic features we propose in Chapter 5, covering this line of work serve as indicators of which methods could be applicable to our work in future directions.

In our work, we investigate the use of social images for concept building and we focus on Flickr groups. Chen and al. [130] are among the first to exploit the visual content of groups. They use Flick group search for a set of 62 concepts and rank the returned groups based on 4 factors related to group popularity. They train dedicated SVM models for concepts and use them independently to recommend tags and groups. Ulges et al [131] investigate the usefulness of groups for photo annotation. They build models for a set of selected groups using all of the images kept after duplicate removal. Then, they induce new annotations for a test image using the tag distribution from the group with the highest predicted probability for that image. However, they mine a relatively limited number of groups (up to 609) and do not aggregate them.

2.1.3 Image collection representation for content based image retrieval (CBIR)

The efficient representation of image collections relies on two main types of structures: partitioning trees [132] and inverted indexes [133]. Partitioning trees are well adapted for an approximate search over dense feature vectors and a number of variations of such structures are discussed in [134]. Classical kd-trees [132] are of limited use when in high-dimensional spaces and approximations were proposed that implement either error bounds [135] (i.e. considering a subspace around the true nearest neighbors) or time

bounds (i.e. the search is stopped after a predefined number of leaves is considered) [136]. The authors of [134] perform a thorough evaluation of different types of tree structures and show that no structure performs best over all evaluation datasets. Depending of the dataset best results are reported with randomized k-d trees and with a variant of a k-means tree. Following [137], a distributed version of k-d trees is proposed in [134] in order to scale-up the search process. Consequent search time reduction is reported in [134], 10^3 - 10^4 acceleration with a precision loss between 5% and 50% compared to exhaustive search. However, the search time is still heavily dependent on the collection size and scaling-up the system requires adding new machines each time the collection grows.

Inverted index structures have been used for Web and text search for many years. It entails mapping each query word to a matching list of documents. The index servers then determine a set of relevant documents by intersecting the hit lists of the individual query words, and they compute a relevance score for each document [138].

An inverted index data structure has been previously proposed for image retrieval. In the standard visual words based inverted indexing structure [51], each visual word is associated with an inverted list, in which the image identification and the frequency of the visual word occurring in the image are stored. Cao et al. [139] use binary counts of spacial bags of visual features based on SIFT descriptors. A more complex approach is detailed in [140]. There, a coding/decoding scheme used for the compression of tree-structured vector quantizer constructed by hierarchical k-means clustering of SURF descriptors.

Jegou et al. [61] proposed an inverted file system, IVFADC, which combines an inverted structure with asymmetric distance computation (ADC). By K-means, IVFADC trains a coarse quantizer of k centroids. Each centroid is associated with an inverted list in the indexing structure. Every descriptor is allocated to the nearest centroid, and the residual vector between the descriptor and its nearest centroid is quantized and encoded by PQ. The descriptor identification information and its codes are stored in the corresponding inverted list according to its nearest centroid. Babenko and Lempitsky [141] proposed an inverted multi-index, which is a multidimensional table based on PQ. Typically, the inverted multi-index partitions descriptors into two sub-vectors. PQ is separately adopted to train two quantizers for the two sub-vectors. The centroid pairs from the two quantizers form the indexing structure of a 2D indexing table. Given a descriptor, the pair of quantization codes by PQ is used as the indices, by which the descriptor is inserted into the corresponding inverted list. For very similar retrieval complexity and pre-processing time, the inverted multi-index achieves a much denser subdivision of the search space compared to the inverted indexing structure from [61], while retaining memory efficiency.

More recent, Zheng et al. [142] proposed a coupled MultiIndex (c-MI) framework to perform feature fusion at indexing level. Basically, complementary features are coupled into a multi-dimensional inverted index. Each dimension of c-MI corresponds to one kind of feature, and the retrieval process votes for images similar in both SIFT feature spaces.

Tavenard et al. [143] introduced a balanced cluster scheme to produce clusters of much more even size. The key idea of this approach is to artificially enlarge the distances from the descriptor to the centroids of the heavily filled clusters so as to shrink and slightly drain the loaded cluster. This is realized by designing a penalization term, where the distance between the descriptor and a centroid is the sum of the Euclidian distance and the penalization term. The more heavily is the cluster filled, the larger the penalization term. Very recently, Liu et al. [144] link previous works dealing with inverted index structures in image retrieval and CNN image descriptors. Instead of projecting each CNN feature vector into a global hashing code, they propose a framework that adapts the BoW model and inverted indexes to global feature indexing. However there is little novelty in their approach, besides the use of CNN features. They simply treat each dimension of the feature vector as corresponding to a virtual concept word and build a dictionary whose size is equal to length of the feature vector.

Our approach of using high level semantic features in an inverted index structure for fast image retrieval is closer to the original textual inverted index than in the previous referenced works. This statement is supported by the fact that, in our case, the keys are concepts conveying semantic meaning. This framework is detailed in Chapter 5.

2.2 Credibility in Information Retrieval

Credibility, as the general concept covering trustworthiness and expertise, but also quality and reliability, is strongly debated in philosophy, psychology, and sociology. Its adoption in computer science is therefore fraught with difficulties. Through this Thesis, we introduce the concept of credibility in a new domain (image sharing platforms), and we propose and analyze multiple credibility estimates. We also explore the uses of credibility in information retrieval.

In this Chapter, we present a detailed study of existing credibility models from different areas of the Web. Nevertheless, the main focus of this Chapter is on research directions in the study of credibility in information retrieval systems. We review here the series of factors that contribute to credibility assessment in human consumers of information, then models used to combine these factors, followed by methods to predict credibility. A smaller section is dedicated to informing users about the credibility learned from the

data. The study then delves into the analysis of credibility in social networks, followed by issues addressing multimedia data. We have attempted to make each of these topics self-sufficient, such that the reader has the option to jump directly to any of these sections.

There is a very rich body of work pertaining to different aspects and interpretations of credibility, particularly for different types of textual content (web sites, blogs, tweets etc.), but also to different modalities (e.g. images, videos). We start the study with an introduction defining basic underlying concepts and placing the concept in the context of other sciences. Following that, we provide a definition of the four components thought to form 'credibility', and consider in detail each of them, with its unique properties and peculiarities. These works serve both as a medium of compiling a unified model for Web credibility that can be applicable in our context, as well as offering clues on which type of Web data can serve as indicators for credibility.

Addressing credibility in the tradition of information retrieval—using benchmarks—is relatively new and the number of available test collections is extremely limited. We present here a set of datasets used for credibility assessment in different domains, as well as the most popular Web resources used to automatically construct test collections. Considering the limited resources available for evaluating credibility in the multimedia domain, we introduce a novel evaluation dataset in Chapter 5.

2.2.1 Credibility Components

In one of the first studies on online information credibility, Fogg and Tseng [145] identify two key components of credibility: *trustworthiness*, which captures the perceived goodness or morality of the source and *expertise*, which relates to the perceived knowledge and skill of the source. Besides these, we also consider the content dimension of credibility, which is linked to that of the source, and includes the concepts of *reliability* and *quality*, as can be seen in Figure 2.1.

The user and content axes of credibility may appear as separate research directions but, as it is a common assumption that a credible source produces credible content and vice-versa, these two axes often intertwine. This relation can be found in studies on credibility, where user profile information is analyzed together with content features [146, 147] or it can be explicitly modeled, such in the case of Bian et al. [148], where they propose a mutual reinforcement framework to simultaneously calculate the quality and reputation scores of multiple sets of entities in a network. Although in general there is a positive correlation between source and content credibility, there are examples from the community question answering domain, where the relationship between user reputation and content quality is not always evident. Users that are highly regarded in

FIGURE 2.1: Aspects of credibility.

the community may provide poor answers, and users with a bad answering history may sometimes provide excellent answers [149].

We revisit this model of Web credibility in the context of user credibility in image sharing platforms in Chapter 5. Next, we will define and briefly detail these four concepts focusing on their relation with credibility. Throughout this Section, we will focus on works and resources related directly to credibility but we will also take into consideration relevant research on the adjacent concepts when they can be linked to credibility.

2.2.1.1 Expertise

Many of the first studies describe the use of expertise finding systems within specific organizations and rely on data sources available within the organization. For example, Expert Seeker [150] was used to identify experts within the NASA organization, relying on a human resource database, an employee performance evaluation system, a skills database, or a project resource management system.

In the social media environment, the number of studies that examine knowledge sharing and expertise increases. Expertise analysis and prediction has been applied on forums [151], online communities [152], blogs [153] and collaborative tagging [154]. A few studies

referred to particular types of social media applications inside the enterprise. Kolari et al. [155] presented an application for expertise location over corporate blogs using the content of the blog posts, their tags, and comments. Amitay et al. [156] presented a unified approach that allowed searching for documents, people, and tags in the enterprise. Data was derived from applications for social bookmarking and blogging, but the two data sources were not compared and the system was evaluated as a whole. Guy et al. [157] focus on comparing a wide variety of enterprise social media applications as data sources for expertise inference.

Within recent works on expertise in online communities, research covering expertise on Community Question Answering stands out. Recently, various approaches have been proposed to automatically find experts in Question answering websites. Jurczyk and Agichtein [158, 159] adopt the HITS algorithm [160] for author ranking. They represent the relationship of *asker* and *answerer* as a social network and calculate each user's hub and authority value. They then rank users according to their authority values. Liu et al. [161] use an expert profile built from the contents of the expert's questions and answers, in order to find experts without considering their reputation and their authority values derived from link analysis but rather from the content of their answers. They recast the problem as an information retrieval problem and use several language models to represent the knowledge of each user. We propose a different study of user profiling for expertise in Question Answering communities in Appendix A.

Regardless of the method of estimation, expertise information is likely to change over time. Rybak et al. [162] introduce a temporal expertise profiling task. This tasks deals with identifying the skills and knowledge of an individual and tracking how they change over time. To be able to capture and distinguish meaningful changes, the authors propose the concept of a hierarchical expertise profile, where topical areas are organized in a taxonomy. Snapshots of hierarchical profiles are then taken at regular time intervals. They propose methods for detecting and characterizing changes in a person's profile, such as, switching the main field of research or narrowing/broadening the topics of research. Contrary to these works, we do not investigate expertise over time. We consider a user to be credible if his or her expertise is constantly reflected through reliable contributions, regardless of the time when they were generated.

In the context of this Thesis, we view expertise either as real life photography expertise or validation received by the community. Indicators of expertise may include clues in the user's description (*e.g.* working for a professional photography institution) or being invited to exclusive curated Flickr groups.

2.2.1.2 Trust

We notice a difficulty in defining trust in general. In computer science, most approaches to credibility strongly emphasize authority, where a trusted source is used to inform an individual's credibility determinations [163]. Trusted sources are used as an indicator for the credibility of a given piece of information.

In fact, many works use the concepts of credibility and trust interchangeably while studying trust in the domain of blogs [164], Wikipedia [165], Twitter [166], or Social Question Answering websites [167]. Others use the notion of trust to identify good quality content and to filter spam [168]. More recently, Toma [169] proposes a framework that identifies cues associated with trustworthiness in Facebook profiles; in this work again, *credible* is used as a direct synonym of *trustworthy*, when referring to cues provided by the friends of a user rather than the user himself or herself, as these are perceived to have less of a motive to embellish or mar a friend's profile.

A distinct area of literature is that on trust in a network environment. If the previously mentioned works identified trustworthiness cues in the data itself, Guha et al. [170] study the problem of propagating trust and distrust among Epinions[1] users, who may assign positive (trust) and negative (distrust) ratings to each other. The authors study ways of combining trust and distrust and observe that, while considering trust as a transitive property makes sense, distrust cannot be considered transitive. Bachi et al. [171] extend the work of trust on Epinions. They propose a global framework for trust inference, able to infer the trust/distrust relationships in complex relational environments in which they view trust identification as a link sign classification problem. In addition to Epinions, they also test their framework on Slashdot[2], where a user can mark another user as friend or foe, and on Wikipedia[3], where the network is extracted from the votes cast by the users in the elections for promoting users to the role of administrator. Ziegler and Lausen [172] also study models for propagation of trust using a spreading activation-inspired model for semantic Web data. They also present a taxonomy of trust metrics and discuss ways of incorporating information about distrust into the rating scores.

Our focus is on image sharing platforms. In this context, we consider *trust* to be related to how a user is perceived by the community. Indicators of trust may include the number of users that have him/her among their contacts, or the comments the user receives for his/her photos.

[1] http://www.epinions.com/
[2] http://slashdot.org/
[3] http://www.wikipedia.org/

2.2.1.3 Quality

Perceptions of quality are closely associated with credibility. Some works identify quality as the super-ordinate concept [173], some view quality and credibility as associated with separate categories [174], and some regard quality as subordinate to credibility [175]. Quality can also be linked to the interest that certain content can raise (i.e. something is of *"quality"* if it is useful/interesting to the audience). Alonso et al. [176] study the problem of identifying uninteresting content in text streams from Twitter. They find that mundane content is not interesting in any context and can be quickly filtered using simple query independent features. Nevertheless, the primary focus when observing the literature on quality centers around stylistic analysis and spam.

Text Quality Analysis When dealing with textual data of any size, ranging from a few characters, such in the case of a Twitter message, to the length of a book, one of the most important features for estimating the credibility of the transmitted message is the quality of the text. This is especially important when there is little or no information about the source of the text or when the truthfulness of the content can not be easily verified.

One encounters considerable amount of work on estimating the quality of text in the field of Automated Essay Scoring (AES), where writings of students are graded by machines on several aspects, including style, accuracy, and soundness. AES systems are typically built as text classification tools, and use a range of properties derived from the text as features. Some of the features employed in the systems are:

- lexical, e.g. word length;
- vocabulary irregularity, e.g. repetitiveness [177] or uncharacteristic co-occurrence [178];
- topicality, e.g. word and phrase frequencies [179];
- punctuation usage patterns;
- the presence of common grammatical errors via predefined templates [180](e.g. subject-verb disagreements).

A specific perspective with regards to text quality is readability. In this case, the difficulty of text is analyzed to determine the minimal age group able to comprehend it. Several measures of text readability have been proposed. Unigram language models were used on short to medium sized texts [181, 182]. Furthermore, various statistical models

were tested for their effectiveness at predicting reading difficulty [183] and support vector machines were used to combine features from traditional reading level measures, statistical language models and automatic parsers to assess reading levels [184]. In addition to lexical and syntactic features, several researchers started to explore discourse level features and examine their usefulness in predicting text readability [185, 186].

Feng et al. [187] compared these types of features and found that part-of-speech features, in particular nouns, have significant predictive power; moreover, that discourse features do not seem to be very useful in building an accurate readability metric. They also observed that among the shallow features, which are used in various traditional readability formulas (e.g. Gunning-Fog Index or SMOG grading [188]), the average sentence length has dominating predictive power over all other lexical or syllable-based features.

Based on an initial classification proposed by Agichtein et al. [149], we identify the following groups of textual features used to reveal quality content:

- *Punctuation*: Poor quality text, particularly of the type found in online sources, is often marked with low conformance to common writing practices. For example, capitalization rules may be ignored, excessive punctuation, particularly repeated ellipsis and question marks, may be used, or spacing may be irregular. Several features that capture the visual quality of the text, attempting to model these irregularities are punctuation, capitalization, and spacing density (percent of all characters), as well as features measuring the character-level entropy of the text.

- *Typos*: A particular form of low quality are misspellings and typos. Additional features quantify the number of spelling mistakes, as well as the number of out-of-vocabulary words. These types or features are found to be useful in several tasks, such as credibility inspired blog retrieval [147] or deriving credibility judgments of Web pages [189].

- *Grammar*: To measure the grammatical quality of the text, several linguistically-oriented features can be used: part-of-speech tags, n-grams or a text's formality score [190]. This captures the level of correctness of the grammar used. For example, some part-of-speech sequences are typical of correctly formed questions.

- *Writing style complexity*: Advancing from the punctuation level to more complex layers of the text, features in this subset quantify the syntactic and semantic complexity of it. These include simple proxies for complexity such as the average number of syllables per word or the entropy of word lengths, as well as more advanced ones such as the readability measures [181, 186].

In Chapter 5, we propose several text quality based user credibility estimators for Flickr users. We look at image tags, titles and description.

Spam as an indicator for bad quality While the dictionary definition of spam[4] still refers exclusively to email, the term has taken a larger meaning in the last decade, referring to all means of undesired, generally commercial, communication.

When referring to Web pages, we can even differentiate between content and link spam:

- *Content spam*: Content spam refers to changes in the content of the pages, for instance by inserting a large number of keywords [191]. Some of the features used for the classification include: the number of words in the text of the page, the number of hyperlinks, the number of words in the title of the pages, the redundancy of the content, etc. Ntoulas et al. [192] show that spam pages of this type can be detected by an automatic classifier with a high accuracy.
- *Link spam*: Link spam may include changes to the link structure of the websites, by creating link farms [193]. A link farm is a densely connected set of pages, created explicitly with the purpose of deceiving a link-based ranking algorithm. Becchetti et al. [194] perform a statistical analysis of a large collection of Web pages, build several automatic Web spam classifiers and propose spam detection techniques which only consider the link structure of Web, regardless of page contents. Andersen et al. [195] propose a variation of PageRank, Robust PageRank, that is designed to filter spam links.

Spam is not limited to Web pages and has been well studied in various applications, including blogs [196], videos [197, 198], Twitter [199, 200], Facebook [201], opinions [202], and of course, e-mail (text based [203] or using multimedia content [204]). Automatic methods for detecting spam are especially useful for exposing sources of weak credibility.

In the context of our work, we do not treat spam as a tagging action with malicious intentions. For Flickr images, we consider bulk tagging (*i.e.* the action of tagging a large set of images with the same tags) to be similar to spam and is the most obvious sign of bad quality tagging.

Benevenuto et al. [205] first approached the problem of detecting spammers on video sharing systems. By using a labeled collection of users manually classified, they applied a hierarchical machine learning approach to differentiate opportunistic users from the non-opportunistic ones in video sharing systems. In a later work, Benevenuto et al. [198]

[4] Merriam-Webster online `http://www.merriam-webster.com/dictionary/spam`

propose an active learning algorithm that reduces the required amount of training data without significant losses in classification effectiveness.

For the credibility work approached in this Thesis, we look at the quality of image tagging and tot the photos themselves. Imposing this restriction for the term *quality*, we consider an image to have good quality tags if they are relevant to the visual content of the image. We note here the difference from *truthfulness*. For example, a user may tag his or her images with the type of camera they were taken with or the date when the photos were taken. While true for the user, these tags serve no purpose for describing the content of the image and cannot be used in a retrieval scenario.

2.2.1.4 Reliability

Reliability commonly refers to something perceived as dependable and consistent in quality [163]. More specifically, text content reliability can be defined as the degree to which the text content is perceived to be true [206]. According to Rieh [207], reliability of the content is a criterion that, following topic relevance, is one of the most influencing aspects that should be considered for assessing the relevance of a Web publication. Connections between the field of credibility analysis and reliability can be found in works dealing with the credibility assessment of blogs [146, 147, 208]. In these works, credibility is applied to multiple concepts besides the reliability measure, and reliability is viewed as a subarea of credibility. Besides being used as a component of credibility, some works place the concept of reliability as the central object of research. Sanz et al. [209] use a combination of information retrieval, machine learning, and NLP corpus annotation techniques for a problem of text content reliability estimation in Web documents and Sondhi et al. [210] propose models to automatically predict reliability of Web pages in the medical domain.

In this Thesis, we have a view on *reliability* that diverges from the works described above. We see reliability as the sustained tagging quality (following the definition presented in the previous Section) of a user's images in time.

2.2.2 Credibility Research Directions

Having clarified our understanding of credibility, we move on to consider the different research areas at the confluence of computer science, information science and credibility. This subsection covers: where do credibility requirements come from and what are the features of the data we can analyze to represent credibility (Section 2.2.2.1); how to predict credibility, or otherwise quantify, based on the features and requirements, the expectation that the user will find the information credible (Section 2.2.2.2); and, finally,

how to inform the user about credibility (Section 2.2.2.3). The following sections will pick up on some of the topics described here.

2.2.2.1 Analysing Credibility

Some of the first impressions on the credibility of a Web page are based on surface credibility which corresponds to the website's appearance: appealing, professional aspect, the website's domain and an important role is played by the website's overall aesthetics [145]. Alsudani and Casey [211] perform a thorough study on the link between aesthetics and credibility. For their survey, 30 people were selected to judge credibility; subjects were of a balanced gender: 15 males and 15 females aged between 18 - 40, all of them being university students. Their study, as well as many others cover a set of typical questions. Table 2.2 shows some of the most common questions in surveys on credibility judgments.

TABLE 2.2: Common questions in credibility surveys.

Visualisations	*Do you think that System X is useful for decision-making?* *Do you think that System X is useful for searching words of mouth?* *Do you think that you can find credible information with System X?*
Implicit Credibility	*Did you find any information that you had expected?* *Is Web page A more credible than Web page B?*

A consistent amount of work has already been dedicated to the study of the influence of source demographics on the perceived credibility of user generated content on the Internet. Flanagin and Metzger [212] analyze the impact of the gender of the source (i.e. not of the assessor/reader, but of the content creator) on the perceived credibility of personal Web pages. They found that men and women had different views of Web site credibility and that each tended to rate opposite-sex Web pages as more credible than same-sex Web sites. A similar study was performed by Armstrong and McAdams [213], who examine the relationship between source credibility and gender. They examine how gender cues influence perceptions of credibility of informational blogs by manipulating the gender descriptors of a blog's authors. They had participants rate the overall perceived credibility of posts and found that male authors were deemed more credible than female authors. What has not been studied in this respect is the influence of cultural background in such perceptions of credibility. Gender and its roles are perceived differently across longitude, latitude, and time [214] and it would be interesting to observe to what extent these perceptions match credibility in the relatively new, information technology world.

However gender is just one of the most obvious and easy to test factors, from the experimental procedure point of view. There has been interest for providing general guidelines for improving the credibility of Web sites based on a more comprehensive set of factors. One such example is the list of 10 guidelines compiled by The Stanford Web Credibility Project[5]. The following suggestions are included in this list:

1. Make it easy to verify the accuracy of the information on your site.
2. Show that there is a real organization behind your site.
3. Highlight the expertise in your organization and in the content and services you provide.
4. Show that honest and trustworthy people stand behind your site.
5. Make it easy to contact you.
6. Design your site so it looks professional (or is appropriate for your purpose).
7. Make your site easy to use—and useful.
8. Update your site's content often (at least show it has been reviewed recently).
9. Use restraint with any promotional content (e.g., ads, offers).
10. Avoid errors of all types, no matter how small they seem.

This and other similar studies are based on theoretical information processing models, like the *Elaboration Likelihood Model* [215] or the earlier *Heuristic-Systematic Model* [216], in the sense that a large component of credibility (in this case referred to as persuasion) is the ability of the user to evaluate the informational content and the intention behind it.

The importance of intention behind the informational content has been shown in a large study based on Web of Trust[6] (WOT) data covering a one year period by Nielek et al. [217]. While they primarily investigate if the websites become more credible over time, the authors also observe that the most credible sites (among 12 categories) are weather forecast sites. They conclude that this is an indicator of the importance of intent in credibility adjudication, since weather forecast is less informationally accurate than news reports of past events, but is seen as unaffected by intentional changes motivated by potentially hidden agendas.

[5] http://credibility.stanford.edu/
[6] https://www.mywot.com/

Building on the Elaboration Likelihood Model's two routes that affect the information readers' attitude towards information (the direct, informational route, and the indirect, information-irrelevant route), Luo et al. [218] perform a study in which they investigate the moderating effect of recommendation source credibility on the causal relationships between informational factors and recommendation credibility. In a second step, the authors also investigate the moderating effect of source credibility on the causal relationship between recommendation credibility and recommendation adoption. This study relates to several of the points in the above list, namely all those related to the ability of the user to identify the source of the information and the ability to assess the credibility of the source independently of the content under current examination.

Making the link between aesthetics and information source, Xu [219] proposes a study in which she explores how two personal profile characteristics, reputation cue and profile picture, influence cognitive trust and affective trust towards the reviewer and perceived review credibility, respectively, in a combinatory manner. The findings of her study showed that the reputation cue (a system generated indicator of reputation) contributed differently from the profile picture to users' trust towards the reviewer: the latter influenced the affective trust alone, while the former influenced both affective and cognitive trust. However, profile pictures are not the only factors used in assessing the personal profile of contributers. For each task, the content consumer uses all information at his or her disposal to assess the user. For instance, in the case of a travel-related task, other self disclosed personal profile information (PPI) would be the reviewer location and travel interest, in addition to the textual content of the review itself [220].

The observation about the profile picture is related to long-standing observations [221] associating physical attractiveness to higher credibility. Physical attractiveness applies in the more general context of website aesthetics and logo design. Lowry et al. [222] analyze the visual content of websites as indicators for credibility, with an emphasis on logo design and propose a 4-point check-list for logo design to enhance credibility, defined by them as a combination of expertise, trustworthiness, and dynamism.

Nevertheless, aesthetics are a more or less important function of the nature of the information to be transmitted. For instance, Endsley et al. [223] study how different factors affect the perception of credibility of crisis information about natural disasters. They find that for crisis information about natural disasters, people tend to trust traditional media channels, such as printed news, and televised news, as opposed to online resources or social media.

The cognitive credibility is supported by the ability of the user to understand the content and to place it in context. One aspect here is accessibility of background information (e.g. references), as a requirement and contributor to credibility. In this sense, Lopes

and Carriço [224] present a study about the influence of accessibility of user interfaces on the credibility of Wikipedia articles. The authors looked at the accessibility quality level of the articles and the external Web pages used as authoritative references. The study has shown that there is a retro-influence of the accessibility of referenced Web pages, which can compromise the overall credibility of Wikipedia. Based on reported results, the authors analyze the article referencing life-cycle and propose a set of improvements that can help increasing the accessibility of references within Wikipedia articles.

Ayeh et al. [225] perform a survey to examine online travelers' perceptions of the credibility of user generated content (UGC) sources and how these perceptions influence attitudes and intentions in the travel planning process. They report mixed results regarding a direct relationship between credibility factors and online travelers' intention to use UGC for travel planning. The direct effect of source expertise on behavioral intention was not supported, while trustworthiness only had a weak effect on behavioral intention. Their findings suggest that trustworthiness and expertise dimensions of source credibility have different importance in affecting attitude and behavioral intention and that trustworthiness is more influential. On the other hand, in a similar study, Xie et al. [226] found perceived source credibility of online reviews to have a significant effect on participants' intention to book a hotel.

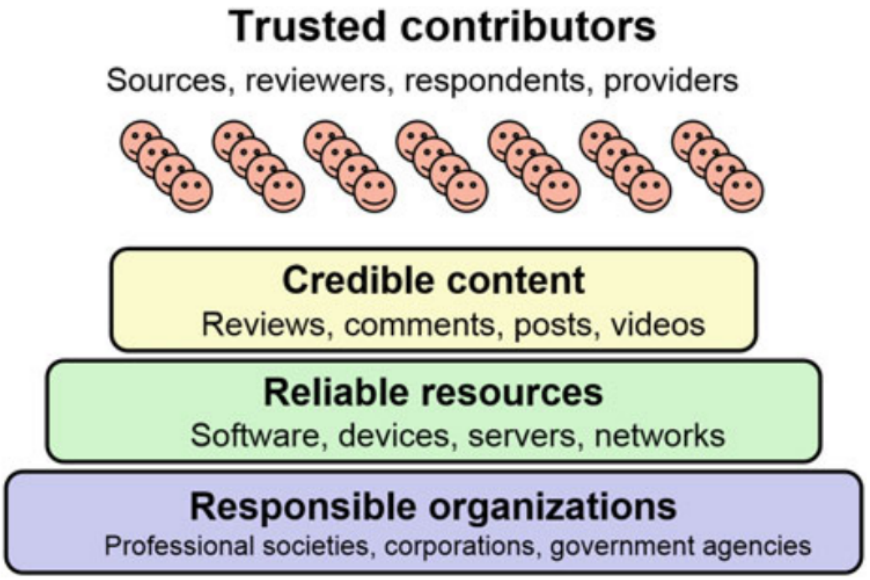


FIGURE 2.2: A framework for analyzing social media communities proposed by Shneiderman [26].

Most recently, Shneiderman [26] starts from all these observations and proposes a framework for analyzing credible communities in social media platforms. As illustrated in Figure 2.2, he hypothesizes that trusted contributors provide credible content that is delivered by reliable resources, guided by responsible organizations. He also points out that

contributors may be misinformed, biased, or malicious, so their content is not credible and physical resources can be undermined.

In fact, the lack of credibility on perceived commercial sales intent has long been documented in the literature [227]. In the Web domain, the presence of intrusive advertisements has also been formally shown to be negative [228], but the relationship is not always simple. Zha and Wu [229] present an experimental study that explores how online disruptive advertisements affect users' information processing, feelings of intrusiveness, and news site's credibility. They find that only if ad content is suspected to co-opt with news production, media credibility suffers. In general, their study (as well as others, like [230]) shows that users filter out Web ads. Especially for the younger generation, the authors hypothesise that the users understand the distinction between advertisements, even intrusive ones, and content.

Among all the factors affecting credibility, two of the following chapters address in more details two sets: social (Section 2.2.3) and multimedia (Section 2.2.4). While not directly to our work, we draw inspiration from this branch of credibility research in developing credibility estimates (Chapter 5).

2.2.2.2 Predicting Credibility

The studies just described relied on extensive user surveys or otherwise crowd-sourced data to understand factors affecting credibility. The following step is using this information to predict what a typical information user will consider credible or not. Developing models able to predict the credibility of the source or content on the Web, without human intervention, is therefore one of the most active research areas in the field of Web credibility. Approaches that have been used for this task include machine learning [36, 210, 231], graphical models [232], link algorithms [168, 233] or game theory [234].

Olteanu et al. [36] test several machine learning algorithms from the *scikit-learn*[7] library (SVMs, decision trees, naïve bayes) for automatically assessing Web page credibility. They first identify a set of features that are relevant for Web credibility assessment, before observing that the models they have compared performed similarly, with Extremely Randomized Trees (ERT) performing slightly better. An important factor for the classification accuracy is the feature selection step. The 37 features they initially considered, as well as those ultimately selected (22), can be grouped in two main categories:

[7] http://www.scikit-learn.org

- *Content features*: refer to features that can be computed either based on the textual content of the Web pages, text-based features or based on the Web page structure, appearance and metadata features.
- *Social features*: include features that reflect the online popularity of a Web page and its link structure.

Jaworski et al. [235] also observe that there is little to no difference in predicting credibility between a simple linear regression method and a neural network model. While the authors do not discuss in great deal the precise nature of the features in [235], their report supports the observations made before by Olteanu et al. [36]. Besides the features introduced in the previous two cited papers, Wawer et al. [236] are also looking for specific content terms that are predictive of credibility. In doing so, they identify expected terms, such as *"energy"*, *"research"*, *"safety"*, *"security"*, *"department"*, *"fed"*, *"gov"*.

Predictors based on content or social features are limited with respect to the transitory nature of credibility. For events rather than general information websites, the information seeking behaviour is rather reactive than proactive: events trigger a cascade of information units which have to be assessed for both informational content and credibility. In such cases, credibility comes as a second step, after an initial phase identifying newsworthiness. Castillo et al. [237] use a supervised learning approach for the task of automatic classification of credible news events. In their approach, a first classifier decides if an information cascade corresponds to a newsworthy event, then, a second classifier decides if this cascade can be considered credible or not. For the credibility classifier, several learning models are tested (Bayesian methods, Logistic Regression, J48, Random Forest, and Meta Learning based on clustering.), with Random Forest, Logistic Regression and Meta Learning performing best and indistinguishably from each other.

Machine learning methods for predicting credibility rely on either user-study data created in the lab, or on crowdsourced data (for instance, from Web of Trust (WOT), or more generally, question answering websites). The latter method can be subjected to credibility attacks by users or methods imitating the behavior of correct users. Machine learning has been used here as well. Liu et al. [238] identify attackers who imitate the behavior of trustworthy experts by copying a system's credibility ratings to quickly build high reputation and then attack other Web content. They use a supervised learning algorithm to predict the credibility of Web content and compare it with a user's rating to estimate whether this user is malicious or not.

Source and content credibility plays an important role in results ranking or re-ranking in information retrieval. The ranking can be obtained from a credibility score specially designed to reflect a particular property of the data [239] or it can result from a learned

value in a supervised manner [149, 240]. Credibility estimation with manual or automatic methods is further developed in the remainder of this Section, as well as in Section 2.2.3 and 2.2.4.

The credibility research direction detailed in this Section is more closely related to our work. Although we work in a multimedia domain, we use a similar credibility features classification as the ones proposed by Olteanu et al. [36]. Also, from a machine learning perspective, we rely on previous studies ([235, 237, 238]) for choosing classification and regression models.

2.2.2.3 Informing About Credibility

Having learned something about the credibility of a website or other information units, the final research direction is how to present this information to the user in such a way that is easy to understand and credible itself.

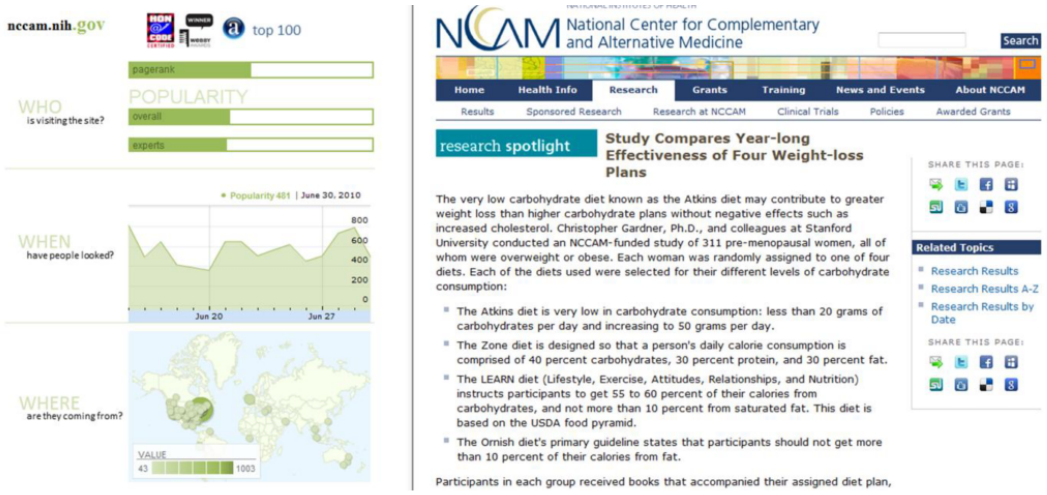


FIGURE 2.3: Example visualization taken from [27].

Schwarz and Morris [27] present visualizations to augment search results and Web pages in order to help people more accurately judge the credibility of online content. They also describe findings from a user study that evaluates their visualizations' effectiveness in increasing credibility assessment accuracy and find that augmenting search results with information about expert user behavior is a particularly effective mean of enhancing a user's credibility judgments.

In Figure 2.3, we show a sample of their visualization. The Web page visualization appears adjacent to the Web page, so that it is visible regardless of scroll positioning. The visualization uses color and font size to draw attention to a page's domain type, and

includes icons to indicate whether a page has received an accredited certification. Horizontal bars indicate the relative value of the current page's PageRank, general popularity, and popularity among experts for the page's topic.

Yamamoto and Tanaka [28] present a system that calculates and provides visualizations of several scores of Web search results on aspects of credibility, predicts a of user's credibility judgment through user's credibility feedback for Web search results, and re-ranks Web search results based on user's predicted credibility model.

FIGURE 2.4: Example visualization taken from [28].

As it can be seen in Figure 2.4, when users run their system on Google's search engine result pages, the system inserts radar charts that illustrate scores of Web search results on each of credibility factors into search results. The users can also re-rank the search results in accordance with their credibility judgment model by double-clicking radar charts of credible Web search results.

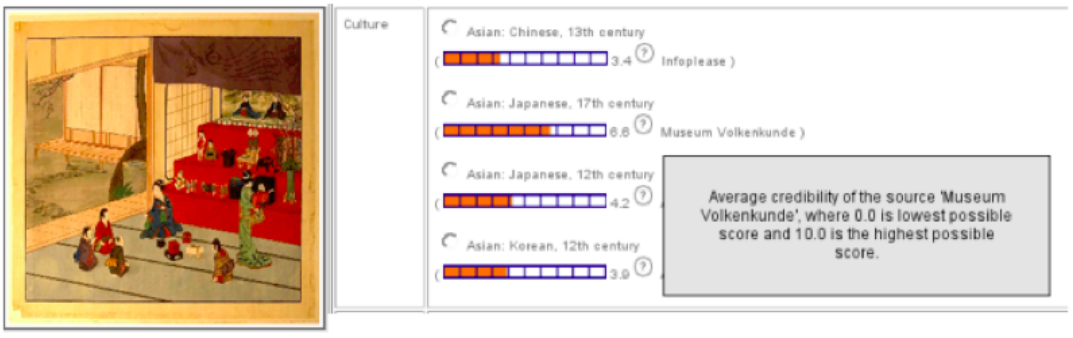


FIGURE 2.5: Example visualization used in [29].

A different, and somehow simpler visualization is presented by Amin et al. [29] (Figure 2.5). Their empirical user study on the effect of displaying credibility ratings of multiple cultural heritage sources (e.g. museum websites, art blogs) on users' search performance investigated whether source credibility has an influence on users' search performance, when they are confronted with only a few information sources or when there are many sources. The results of their online interactive study show that by presenting the source credibility information explicitly, people's confidence in their selection of information significantly increases, even though it does not necessarily make search more time efficient.

Another visualization possibility is to show the trend of opinions and articles on news sites. Kawai et al. [241] make the assumption that if users know the trend of the news site, they can evaluate the credibility of each news topic. Their system detects and uses the sentiment emerging from each news article (i.e. positive/negative sentiment) to resolve the trend of websites. This trend is extracted as average sentiment scores of the news articles that were written concerning a topic in each website.

The alternative to visual displays such as those just described, is to provide the user with the necessary textual information, to enable him or her to see a variety of facts before making a judgment on credibility. For instance, Murakami et al. [242] introduce Statement Map, a project designed to help users navigate information on the Internet and come to informed opinions on topics of interest. The proposed system mines the Web for a variety of viewpoints and presents them to users together with supporting evidence in a way that makes it clear how the viewpoints are related. The authors discuss the need to address issues of information credibility on the Internet, outline the development of Statement Map generators for Japanese and English and detail the technical issues that are being addressed. While this is a very exciting research direction, the authors do not evaluate the results of their method.

In general, the impact of methods designed to help users judge the credibility of Internet content is usually evaluated in a quantitative fashion by conducting focused surveys (online [243] or in person [211]). Akamine et al. [189] asked thirty participants to use their system and answer a questionnaire. The participants were from a wide range of ages and and regular Internet users. The participants were asked to analyze Web pages grouped on four topics with both their system and Google and to answer some multiple questions on a five-level Likert scale.

However, informing about credibility is not always and necessarily based on automatically computed indicators. For instance, in collaborative epistemological resources such as Wikipedia, it is generally the editors who, upon reviewing existing article, introduced credibility indicators such as *"citation needed"*, *"verification needed"*, or *"unreliable source"*

[224]. Additionally, the crowd can also be used, and is in fact currently in commercial use under the Web of Trust model, where, upon installing a browser plugin, each link on a website is accompanied by a colored logo from green to red indicating the crowd-reputation of the website on the other side of the link.

Providing all these indicators, be they automatically calculated as aggregations of credibility aspects, or simply visual cues to known credibility factors (e.g. colored stars based on average reviews), is of course not guaranteed to trigger a specific behavior in users. For instance, Flanagin et al. [244] conducted a large survey and a focused experiment to assess how individuals perceive the credibility of online commercial information, particularly as compared to information available through more traditional channels, and to evaluate the specific aspects of rating information that affect people's attitudes toward e-commerce. The results of this survey show that consumers rely heavily on Web-based information as compared to other channels, and that ratings information is critical in the evaluation of the credibility of online commercial information. The authors conclude that experimental results indicate that ratings are positively associated with perceptions of product quality and purchase intention, but that people attend to average product ratings, but not to the number of ratings or to the combination of the average and the number of ratings together. Following this direction, Rafalak et al. [245] propose a study aimed at identifying various determinants of credibility evaluations. They had 2046 adult participants evaluate credibility of websites with diversified trustworthiness reference index and they focused on psychological factors that lead to the characteristic positive bias observed in many working social feedback systems on the Internet. They find that the level of trust and risk taking are good measures to be included in research dedicated to evaluating websites' credibility and conclude that using the *need for cognition scale* (i.e. one of the scales investigated in their paper) in research connected with evaluating websites' credibility is questionable. This statement is supported by their findings, which show that results obtained in this scale do not differentiate people having tendency to overestimate and underestimate a website's credibility.

Although we do not address directly the subject of credibility inspired visualizations, the results presented in Chapter 5 can serve as a value used for designing a visual credibility indicator for Flickr users. Also, in Chapter 6.2, we use credibility to rerank a list of results in an image retrieval system. The same values used for reranking can serve as a basis for visual credibility clues, similar to those used for Web search [29].

2.2.3 Credibility in Social Networks

Social networks are now an important information source and have been shown to influence even major societal events, such as governmental elections [246, 247]. Johnson and Kaye [248] perform a survey in order to study the degree to which politically interested online users view social network websites as credible and show that, in the case of political campaigns, users find blogs more credible than online newspapers; but, at the same time, they find Facebook and similar services less credible as a whole.

Ranking social media users on their credibility is one approach used to measure the credibility of the given piece of information. Sometimes, indicators of user credibility are explicitly embedded in the website. Twitter, for example, has a set of verified accounts that are accompanied by a badge. This helps users discover high-quality sources of information and trust, insofar as a legitimate source is authoring the account's tweets. Although these initiatives are helpful, social media websites are not able to verify all their users. Moreover, many users would prefer to remain unknown, and it is expected that the majority of users in social media are unverified.

Abbasi and Liu [249] provide a first overview of credibility in social media. They find that there are some works that use link based information (e.g., PageRank and HITS) to rank the users and evaluate the content based on the source's rank. For instance, Jurczyk and Agichtein [158] use HITS to rank users and find experts and high quality answers in the question answering communities. Using the number of in-links (e.g. the number of friends on Facebook or number of followers on Twitter) is a well-accepted feature for measuring the importance or influence of users. Cha et al. [250] use three approaches (in-degree, re-tweet, and mention) to measure users importance in Twitter. Their study shows that although in-degree measures the popularity of a user, it does not necessarily reflect the importance of the user.

In the following sections, we give an overview of general approaches used for assessing credibility in social networks. We group the link based methods in one section and present in the last two sections works on credibility focused on Twitter and question answering platforms, respectively.

2.2.3.1 Global Approaches

To allow for a better presentation of online reviews to users, O'Mahony and Smyth [251] try to determine the helpfulness of reviews. Their features are divided in reputation features, content features, social features, and sentiment features. A follow-up work also includes readability features [252].

Abbasi and Liu [249] study the situations in which the credibility of the content or the credibility of the user cannot be assessed based on the user's profile. They propose an user clustering algorithm that analyses social media users' online behaviour to measure their credibility.

Edwards et al. [253] focus on a popular influence indicator platform and analyze *Klout.com*, a Website that proposes a popular indicator of a user's online influence. The authors propose a study that has the goal to determine whether and to what degree a Klout score can influence perceptions of credibility. They found that the mock Twitter page with a high Klout score was perceived as higher in terms of credibility compared with the identical mock Twitter page with a moderate or low Klout score.

Yaakop et al. [254] examine the online factors that influence consumers' perceptions and attitudes towards advertising on Facebook. A total of 350 respondents participated in the study. Their results suggest that there are three online factors that significantly influence consumers' attitudes towards advertising on Facebook: perceived interactivity, advertising avoidance and privacy. Contrary to foundings reported for Web pages, they state that credibility was not a significant factor in predicting consumer' attitudes towards advertising on Facebook.

2.2.3.2 Link Methods

In social network analysis, link-based methods are one of the most used approaches. In particular, link-based ranking algorithms that were successful in estimating the quality of Web pages have been applied in this context. Two of the most prominent link-based ranking algorithms are PageRank [255] and HITS [160]. ExpertiseRank [152] corresponds to PageRank over the transposed graph. For example, in a question answering website, a score is propagated from the person receiving the answer to the person giving the answer.

The HITS algorithm was applied over the same type of graph [158, 256] and it was shown to produce good results in finding experts and/or good answers. Jurczyk and Agichtein [159] demonstrate that HITS is a promising approach, as the obtained authority score is better correlated with the number of votes that the items receive than simply counting the number of answers the answerer has given in the past. Dom et al. [257] studied the performance of several link-based algorithms to rank people by expertise on a network of e-mail exchanges, testing on both real and synthetic data, and showed that on real data ExpertiseRank outperforms HITS.

Similar to the works presented above, we test both PageRank and HITS as possible credibility estimators for Flickr users using the Flickr contacts network in Chapter 5.

2.2.3.3 Twitter

Similar to blog credibility, research dealing with credibility in the microblogging environment, represented by the Twitter platform, targets one or several of the credibility dimensions mentioned in Section 2.2.1. Works that treat credibility as a central theme, in general [231, 243, 258–261] , by topic [262] or event credibility [263, 264], are accompanied by those on expertise [265, 266], trust [166, 267, 268], influence [233, 250] and spam [199, 269–271]

Credibility-inspired indicators have been successfully applied to post finding in microblogs [272]. Besides translating indicators from blog credibility to the new environment, the authors also introduced platform-specific indicators like followers, retweets, and recency. For the task of exploring trending topics on Twitter, Castillo et al. [258] use a similar set of indicators to assess the credibility of tweets and use human assessments to test their approach.

Sikdar et al. [273] propose a methodology for developing studies that introduce methods to make credible data more useful to the research community. In this scope, they offer a couple of guidelines. Firstly, they point out the importance of the underlying ground truth values of credibility, that has to be reliable, as well as the specific constructs used to define credibility, that must be carefully described. By proposing these guidelines, they offer an important theoretical framework for future efforts in credibility ground truth construction. Secondly, they consider that the underlying network context must be quantified and documented. To illustrate these two points, the authors conduct a unique credibility study of two different data sets on the same topic, but with different network characteristics. They also conduct two different user surveys, and construct two additional indicators of credibility based on retweet behavior. In a follow-up work, Sikdar et al. [274] propose two methods for identifying credible information in Twitter. The first one is based on machine learning and attempts to find a predictive model based on network features. Their method is geared towards assessing the credibility of messages. The second method is based on a maximum likelihood formulation and attempts to find messages that are corroborated by independent and reliable sources.

In a typical research setting of analyzing credibility factors through users studies, as presented in Section 2.2.2.1, Westerman et al. [275] examine how pieces of information available in social media impact perceptions of source credibility. Participants in the study were asked to view 1 of 3 mock Twitter pages that varied in the recency with which tweets were posted and then to report on their perceived source credibility of the page owner. In a similar work, Aladhadh et al. [276] investigate how certain features affect user perceptions of the credibility of tweets. Using a crowdsourcing experiment,

they found that users' perception of the credibility of tweets is impacted more by some features than by others, most noticeable being the fact that displaying the location of certain types of tweets causes users viewing these tweets to perceive them as more credible.

Shariff et al. [277] also examine user perception of credibility, with a focus on news related tweets. They conduct a user study on a crowdsourcing platform to judge the credibility of such tweets. By analyzing user judgments and comments, they find that eight features, including some that can not be automatically identified from tweets, are perceived by users as important for judging information credibility. Moreover, they find that distinct features like the presence of links in tweets, display name and user belief consistently lead users to judge tweets as credible and that users can not consistently judge or even misjudge the credibility for some tweets on politics news.

Kostagiolas et al. [278] have conducted a recent study in which they consider a simple trust model, according to which they assume that perceived trust is a direct antecedent of perceived credibility. They evaluate whether work-related or personal motivating factors influence the relation between perceived credibility and trust toward institutional information sources and how each factor affects this relation. Their findings suggest that work-related factors have a higher impact on the relation between credibility and trust than personal motivation factors, while they are stressing the important role of hospital libraries as a dissemination point for government-sponsored information resources.

2.2.3.4 Community Question Answering (CQA)

Fast access to relevant information is particularly important when complex information needs, such as learning about a new topic or solving a specific problem, are expressed. When such needs occur, people often consult relevant Web communities (forums, Q&A websites etc.) which gather contributions from a large array of users with different levels of expertise. However, unlike the works focused on Twitter, where the term credibility is often directly used, in most of the papers analyzing question answering communities, we mostly encounter 3 out of 4 credibility components (expertise, quality and trust) that we proposed at the beginning of this survey, in Figure 2.1.

According to Su et al. [279], the quality of answers in question answering portals is good on average, but the quality of specific answers varies significantly. Jeon et al. [280] extracted a set of features from a sample of answers in Naver, a Korean question answering portal similar to Yahoo! Answers. They built a model for answer quality based on features derived from the particular answer being analyzed, such as answer

length, number of points received, etc., as well as user features, such as fraction of best answers, number of answers given, etc.

The quality, accuracy, and comprehensiveness of the content in the CQA archives varies drastically, and a large portion of the content is not useful for answering user queries. The reputation and expertise of the contributors can provide crucial indicators for the quality and the reliability of the content. The reputation of the contributor could also be a valuable factor for ranking search results from CQA repositories, as well as for improving the system interface and incentive mechanisms.

Existing methods for estimating content quality in CQA may use supervised classification methods [149] or focus on the network properties of the CQA without considering the actual content of the information exchanged [152]. Su et al. [281] try to detect text trustworthiness by incorporating evidence of phrases denoting a high confidence in their feature set.

Bouguessa et al. [282] argue that an empirical distinction between expert and non-expert contributors to a CQA hampers the overall quality of expert detection. Using a graph based view of the community, they introduce a principled model for authority scores that is based on a mixture of gamma distributions. Then they show that this model is well fitted for the problem posed. Liu et al. [283] report that adding domain expertise and user reputation to graph-based features improves expert identification in CQA. However, they only exploit votes given to a user's answers to derive domain expertise and reputation and disregard other relevant user data, such as demographic factors or completeness of self-description. Bian et al. [284] discusses the shortcomings of supervised expert detection approaches (e.g. the availability of a large set of labeled data) and introduce a semi-supervised method based on coupled mutual reinforcement. Their framework is capable of finding high-quality answers, questions as well as experts by combining a comprehensive array of question, answer and user features. Liu and Agichtein [285] analyze answerer behavior to determine when and how answers are generated. They confirm that users have daily and weekly periodicities but also point out that there are bursty patterns of activity. Equally interesting, users have favorite categories in which they provide answers but the choice of the questions they answer is mostly determined by their rank in the list of available questions. In a related study, Pal and Konstan [286] show that expert and non-expert CQA contributors can be differentiated based on a selection bias that is stable over time. Experts tend to choose questions for which they have a chance to make a valuable contribution.

Furthering the research dealing with expertise in QA communities, in Appendix A, we focus on the automatic assessment of CQA activity. We investigate if user profile information contain useful hints for expertise discovery, if topic discovery in past contributions

can be used to predict the quality of new answers and, finally, if profile and activity data can be effectively combined in an automatic answer reranking scenario. We first present features that contribute to discriminating expert users. Then we discuss two application scenarios: quality prediction for newly arrived answers and an automatic answer reranking. Thorough evaluations using Stackoverflow content are proposed for both scenarios. The evaluation results show that an analysis of user profiles highlights interesting clues for expertise and the methods introduced in this work significantly outperform appropriate baselines.

2.2.4 Multimedia Credibility

As we have seen in preceding sections, the credibility of textual content was already thoroughly studied from different angles. In contrast, we were able to find only a limited quantity of studies dealing with the credibility of multimedia content. We describe them in this Section. Most of existing works deal either with video or audio content and there is little prior work concerning the combination of textual and visual features for automatic credibility estimation. Very recently, works dealing with credibility in image sharing platforms, such as Flickr, started to emerge, mainly in the context of image retrieval. Estimating the credibility of the source has been proven to be beneficial for the performance of an image retrieval system [287]. This has been also confirmed by the introduction of user credibility in the 2014 MediaEval Retrieving Diverse Social Images Benchmarking Initiative [20], where some of the participating teams [288, 289] have improved the relevance and diversity of an image retrieval system using user credibility estimators. Although these works treat credibility in a multimedia domain, there are only few initiatives in this direction, almost exclusively centered around the dataset introduced for the MediaEval Benchmark [290]. One of the main contributions of this Thesis is to offer a detailed analysis of credibility in the multimedia domain by proposing the study of credibility in image sharing platforms (Chapter 5) and using credibility in the image retrieval domain (Chapter 6). In the following, we focus on credibility works carried in different multimedia mediums: video (Section 2.2.4.1) and audio (Section 2.2.4.2).

2.2.4.1 Video Content Credibility Analysis

Video content is afflicted by the same credibility incertitude as any other type of user generated content. Most of the time, there is not any certain information regarding where did it come from and who produced it or what kind of expertise has the person who produced that resource. Similar to the Health On the Net initiative for medical websites, a couple of organizations proposing to regulate the information spread through

online media content, including videos, have appeared. These are sometimes referred as *media watchdogs* and include websites such as Politifact[8] and FactCheck[9]. They address issues of information quality by combing through the media and engaging in fact-checking of news and other media reports. While most methods of watchdogging are time consuming, another method of coping with information quality includes harnessing social information processing systems [291] which seek to filter information and identify quality by aggregating the recommendations and ratings of many users through passive (e.g. through usage) or active (e.g. through voting or active rating) metrics of recommendation.

Visualizations for Video Credibility Assessment Diakopoulos and his colleagues at the University of Maryland propose solutions that help users judge the quality of a video posted on different online video sharing platforms and that provide hints for the credibility of the video in terms of context and information content.

FIGURE 2.6: Example extracted from the Videolyzer presentation video.

In their first work, Diakopoulos et al. [292] build and study the usefulness of a tool, Videolyzer[10], designed to aid political bloggers and journalists in the activity of watchdog journalism, the process of searching though and evaluating the truthfulness of claims in the media. Videolyzer follows a video quality annotation scheme described in [293] that allows users to collectively analyze the quality of online political videos and then aggregate and share these analyses with others. Users can assess aspects of quality in

[8] http://www.politifact.com/
[9] http://www.factcheck.org/
[10] http://www.nickdiakopoulos.com/projects/videolyzer-information-quality-analysis-for-videos/

the video, its transcript and annotations including bias, accuracy, and relevancy that can then be backed up with sources and reasons. We provide a sample extracted from the Videolyzer presentation video in Figure 2.6.

Diakopoulos and Essa [30] also propose a video player augmented with simple visuals that indicate aggregated activity levels and polarity of evaluations (i.e. positive / negative) shown in-line with videos as they play. Users are able to interact with the visualization for the details of the evaluations including tags, sources, and comments. In Figure 2.7, there is an example of the video annotation system. Layered over the bottom of the video, the graphic depicts the activity and polarity of annotations as a stacked line graph which is time-aligned to the timeline of the video. Negative annotations are red, positive are green, and neutral are gray. As the video plays, the timeline thumb advances and intersects the graph to show the relevant part of the graph. Interaction with the graph reveals two additional layers of information, which are shown in panels that pop up, such as the number of contributors to these annotations, shown in text. Other annotations may be the text of a comment, a tag, or a link to a supporting source. The user can scroll through and read the entire message there. All of these visuals roll-up from the bottom of the video and are designed to be tightly integrated with watching the video itself.

FIGURE 2.7: Example of the video annotation system taken from [30].

In order to understand the influence of this visualization on casual video consumption, they evaluate its impact on the credibility of the information presented in the video as compared to a control presentation of the video. They find that for the negatively annotated videos, the graphic on credibility ratings has a stronger effect on users who engaged the graphic more.

Credibility Prediction in Video Sharing Platforms Besides developing novel visualization methods to highlight credible content in videos, special attention has been given in building automatic methods to predict the quality of the content posted in online

video sharing platforms. Following the trend of the latest research on Web credibility that focuses on the credibility of the users, Benevenuto et al. [198] aim to detect users who disseminate video pollution, instead of classifying the content itself. They use features that capture the feedback of users with respect to each other or to their contributions to the system (e.g., number of views received), exploiting their interactions through video responses. A machine learning approach is devised that explores the characteristics of manually classified users to create models able to identify spammers and promoters on YouTube. In a complementary approach, O'Callaghan et al. [294] take advantage of the network of video propagation in YouTube and apply network analysis methods to identify spam campaigns. Content based classification imply combining multiple features extracted from textual descriptions of the video such as tags, title, textual description and from the video content itself. Boll [295] finds that these types of features are often robust for the typically low quality of user-generated videos.

We have included in the quality component of credibility works that deal with spam. This behavior is not restricted to email or Web pages, but also multimedia content. Bulakh et al. [296] collect a sample of over 3,300 fraudulently promoted YouTube videos and 500 bot profiles that promote them. They characterize fraudulent videos and profiles and train supervised machine learning classifiers that can successfully differentiate fraudulent videos and profiles from legitimate ones. They find that an average fraud video has shorter and fewer comments but is rated higher (4.6 on a 5-point scale when an average legitimate video is rated only at 3.6). Also, the profiles which promote the fraudulent videos, have distinct characteristics: they are relatively new in the system but more active than legitimate profiles, they are more active in viewing and interacting with videos and rarely upload any videos.

Visual and Textual Content Correlation for Prediction Xu et al. [297] aim to help users filter multimedia news by targeting credible content. They propose methods to evaluate multimedia news by comparing visual descriptions and textual descriptions respectively as well as considering the relationship between them. Also, they provide to the users results easy to be understood by ranking the multimedia news in the event ordering by their relative credibility scores. They focus their analysis on multimedia news consisting of video clips and their surrounding texts and compute the credibility score of each multimedia news by considering both visual and textual parts. They introduce a Material-Opinion model to compare any two of the multimedia news reporting the same event. The credibility score of a video news item consists of material and opinion credibility scores:

- *Material credibility score* is computed based on the idea that high credible material should be used in most items and they support similar opinions.
- *Opinion credibility score* is based on the idea that high credible opinion should be claimed in many news items by using different materials.

They use the stakeholder model representing the contents for comparing materials and opinions, respectively. The model is described in detail in [298]. Stakeholders are the important entities in the event, whose descriptions are supposed to be the most valuable parts for the comparison. To evaluate their method, they use user credibility ratings (from 1 to 5) of the news items. They find that the credibility-oriented ranking of the multimedia news correlates with the user ratings.

Exploiting the same idea of deriving credibility scores from visual and textual associations on the Web, Yamamoto and Tanaka [299] built ImageAlert, a system that focuses on the credibility of text-image pairs and propose a bipartite graph model for analyzing the credibility of text-image pairs on the Web, in which one set of nodes corresponds to a set of text data, and the other corresponds to a set of images. Each text-image pair is represented by an edge. They introduce the notion of supportive relationships among edges in the bipartite graph model and postulate that the more supportive text-image pairs a target text-image pair has, the more credible it is.

2.2.4.2 Credibility of Online Audio Content

Tsagkias et al. [300] present an ample study on the credibility of podcasts. They describe PodCred, a framework that consists of a list of indicators that encode factors influencing listener perceptions of the credibility and quality of podcasts. The work is performed in an information science perspective and the authors consider credibility to be a perceived characteristic of media and media sources that contributes to relevance judgments, as indicated in [301]. They incorporate quality by using an extended notion of credibility that is adapted for the purposes of the podosphere. In the context of the podosphere, similar to the works of Weerkamp et al. [147, 239] on the credibility in the blogosphere, other components contributing to user perceptions of credibility, such as expertise and trustworthiness [302] are used. Users prefer podcasts published by podcasters with expertise, i.e., who are knowledgeable about the subject, and who are trustworthy, i.e., they are reliable sources of information and they have no particular motivation to deceive listeners. Tsagkias et al. [300] offer an in-depth analysis of the features used for predicting podcast credibility. We present a sample of each type of features used by them:

- *Podcast Content*: spoken content (e.g. appearance of on-topic guests, participation of multiple hosts, use of field reports, contains encyclopedic/factual information etc.) and content consistency (e.g. podcast maintains its topical focus across episodes, consistency of episode structure, presence/reliability of inter-episode references, episodes are published regularly, etc.);

- *Podcaster*: podcaster speech (e.g. fluency/lack of hesitations, speech rate, articulation/diction, accent), podcaster style (e.g. use of conversational style, use of complex sentence structure, podcaster shares personal details, use of broad, creative vocabulary, etc.), podcaster profile (e.g. podcaster scene name, podcaster credentials, podcaster affiliation, podcaster widely known outside the podosphere);

- *Podcast context*: podcaster/listener interaction (e.g. podcaster addresses listeners directly, podcast episodes receive many comments, podcaster responds to comments and requests, podcast page or metadata contains links to related material, podcast has a forum) and real world context (e.g. podcast is a republished radio broadcast, it makes reference to current events, podcast has a store, presence of advertisements, etc.);

- *Technical execution*: production (e.g. signature intro/opening jingle, background music, editing effects, studio quality recording/no unintended background noise), packaging (e.g. feed-level metadata present/complete/accurate, episode-level metadata present/complete/accurate, ID3 tags used, audio available in high quality or multiple qualities, etc.), distribution (e.g. simple domain name, distributed via distribution platform, podcast has portal or homepage, reliable downloading).

Although some ideas are taken from studies of blog credibility, a clear difference between blogs and podcasts is that the core of a podcast is its audio content. For this reason audio and speech characteristics are taken into account when analyzing podcasts. Following the classic separation of source and content credibility, as detailed by Rieh and Belkin [174], message credibility and source credibility overlap to a certain degree and, in the PodCred framework, it can also be seen that certain podcast content indicators could be argued to also be important podcaster credibility indicators. As a direct application of the framework, the authors indicate Podteller[11], an application that computes the probability of a podcast to become popular in its category.

[11] http://zookma.science.uva.nl/podteller/

2.2.5 Credibility Evaluation Datasets

A number of manually validated ground truth credibility evaluation datasets are readily available and we list them in Section 2.2.5.1. They cover mostly textual data and website metadata, but we also describe a recently introduced dataset for credibility evaluation in the multimedia domain. In Section 2.2.5.2, we present evaluation collections that were gathered from Web data with no or minimum human intervention.

2.2.5.1 Manually Built Datasets

In Table 2.3, we present freely available datasets that are annotated with credibility judgments or were used in credibility related research with little or no alteration.

The Morris Web Credibility corpus contains a dataset of 1,000 URLs that have been manually rated for credibility on a five-point Likert scale. A score of 1 corresponds to "very non-credible", and 5 to "very credible". The URL and ratings list are available for download, as well as the page contents as cached at the time of rating. Moreover, additional expert ratings for the 21 pages used in the experiment described in [27] are available (expert raters were two medical doctors, two banking and investment professionals, and two presidential political campaign volunteers).

The MPI-SWS[12] Twitter dataset contains 54,981,152 user accounts that were in use in August 2009 and 1,963,263,821 social (follow) links. The almost 55 million users are connected to each other by 1.9 billion follow links. This is based on the snapshot of the Twitter network topology in August 2009. The follow link data does not contain information about when each link was formed. The dataset also contains 1,755,925,520 tweets. For each of the 55 million users, information about all tweets ever posted by the user since the launch of the Twitter service was gathered. The tweet data contains information about the time each tweet was posted.

Table 2.3: Datasets used in credibility evaluations.

Dataset	Domain	Usage
MPI-SWS	Twitter	Influece detection [250], Spam detection [199]
Morris Web Credibility	Web pages	Credibility [27]
TREC Blog06	Blogs	Credibility based ranking [147, 239]
Div150Cred	Flickr images	Landmark image retrieval and diversification [288, 289]

[12] http://twitter.mpi-sws.org/

TREC Blog06 corpus [303] has been constructed by monitoring around 100,000 blog feeds for a period of 11 weeks in early 2006, downloading all posts created in this period. For each link (HTML page containing one blog post) the feed id is registered.

Div150Cred [290] represents a specially designed dataset that addresses the estimation of user tagging credibility and stems from the 2014 Retrieving Diverse Social Images Task at the MediaEval Benchmarking Initiative [20]. It provides Flickr photo information (the date the photo was taken, tags, user's id and photo title, the number of times the photo has been displayed, URL link of the photo location, GPS coordinates) for about around 300 locations and 685 different users. Each user is assigned a manual credibility score which is determined as the average relevance score of all the user's photos. To obtain these scores, only 50 157 manual annotations are used (on average 73 photos per user).

We propose a novel dataset (Multi-Topic Tagging Credibility Dataset), designed with the goal of analyzing user credibility for a diversified set of topics in Chapter 5.3.

2.2.5.2 Automatically Built Datasets

Building representative ground truth corpora with human annotators is a costly and time consuming process. It is common to derive evaluation corpora from existing resources (e.g. online communities, forums, social networks, etc.) with minimum processing or intervention. Some of the most used resources for credibility related studies are the following:

- *Epinions*: Epinions is a Web site where users can write reviews about products and assign them a rating. It also allows the users to express their *Web of Trust*, representing users whose reviews and ratings they have consistently found to be valuable and their *Block list*, a list of authors whose reviews they find offensive, inaccurate, or in general not valuable. Works that use corpora built from Epinions data generally study trust propagation [170, 171, 304, 305].
- *Wikipedia*: Wikipedia is the most popular source of encyclopedic information. It was used for many studies, concerning mostly data quality [306, 307], but also trust [165, 171, 308] and credibility perceptions [309].
- *Yahoo! Answers*: Yahoo! Answers[13] is one of the largest question answering communities, with more than 1 billion posted answers. Most works on Yahoo! Answers derive their corpus by using the community votes over the answers as quality indicators. Research using Yahoo! Answers revolves around quality [310] , expertise [311] and trust [167].

[13] http://answers.yahoo.com/

- *StackOverflow*: Stackoverflow[14] is one of the most active and popular CQA platforms that covers a wide area of computer science topics. Similar to Yahoo! Answers, user ratings of answers and questions are used as ground truth quality scores. Works using StackOverflow generally cover the topic of expertise [312, 313].

- *Websites white/black lists*: The Health on Net Foundation (HON) and Quackwatch[15] rate websites based on how credible they believe the website is. In some works, these lists are used as positive and negative examples for testing different automatic methods for estimating credibility [210, 314].

[14] http://stackoverflow.com/
[15] http://www.quackwatch.com/

Chapter 3

Large scale visual concept modeling

In this chapter, we propose a scalable image classification framework that exploits binary linear classifiers. To implement this framework, we compare two data sources: a large manually annotated image dataset (i.e. ImageNet) and Flickr groups. Since the second resource is collected from Web images, a key methodological part of the work details methods that reduce the noise inherent to the collection. We also provide a preliminary evaluation of individual visual models built from the two resources. We investigate the influence of the number of negative training instances on the prediction performance and the training time. The obtained classification framework is subsequently exploited in the following chapters.

3.1 Motivation

As predicted a few years ago [10], research in visual and multimedia recognition has strongly benefited from the availability of manually labeled large-scale image and video collections. In conjunction with theoretical advances [315] and quite cheap and efficient hardware, these collections allow the emergence of visual recognition based on convolutional neural networks (CNN) and the entrance in the era of *deep learning*. For instance the ImageNet representation [12] of nearly 22,000 concepts with approximately 14 million images according to a hierarchy of concepts was thoroughly exploited to learn powerful image representations and led to a new state of the art in image classification [13].

In parallel to these mainstream *bottom-up* approaches, several works adopted a *top-down* scheme to design semantically grounded image features. Given the availability of large-scale image datasets, [3, 22] argued that a representation based on the outputs of a bench of base classifiers would offer a rich, high level description of images that is close to the human understanding, and also allow cross-modal (text-image) retrieval. Moreover, they can benefit from the advances of the *bottom-up* works that propose better mid-level features in order to improve the base classifiers. Another advantage of these representations is that they are scalable in terms of number of classes recognized in order to cope with a wide variety of content.

These approaches are very promising but raise new problems, concerning in particular the availability of the underlying resources. Manually labeled datasets are the result of sustained effort provided by motivated communities of researchers [14], eventually supplemented with crowdsourcing [10], [12]. An important limitation of this approach is that manual annotation is a repetitive task and annotators tend to become demotivated. In addition, when conducted on a large scale, crowdsourcing has a non-negligible financial cost and dedicated funding is difficult to obtain. A promising way to circumvent the lack of annotated data is to use images shared on online social networks (OSNs), such as Flickr. An advantage of this type of resource compared to to manually created collections is that data are annotated by a community of users motivated to make their content accessible [15]. The main drawback of user contributed collections is that a part of images annotations is not directly related to the visual content [316].

In this Thesis, we also explore a novel use of individual visual concept classifiers. We are interested in investigating the link between user generated tags and the visual content of an image as a measure of tagging quality. Given that in an image sharing platform, such as Flick, the vocabulary of tags covers a very large number of concepts, Flickr images are a viable alternative for training visual concept classifiers. We therefore need to shift our focus from manually labeled resources towards exploiting user generated content for building concept models. Both using manual resources, where the concepts are well defined in advance, and Web data together with a noise reduction technique lead to effective building blocks for visual concept learning. The concepts resulted from Flickr images are semantically closer to the real world user tagging behavior. We illustrate the difference between the two type of resources in terms of semantic coverage in Section 3.5.1.

3.2 Image representation

3.2.1 Convolutional neural networks (CNN) image descriptors

Convolutional neural networks have recently shown impressive image classification performances in the large-scale visual recognition challenge ILSVRC [63] and have continued to gain interest in the computer vision community by reaching state-of-the-art results in multiple image and video recognition tasks [79, 317]. Compared to traditional low-level features such as Fisher Vector [64], the use of CNN brought down the ILSVRC error rate from 0.26 to 0.15 in 2012, 0.11 in 2013 [14, 63] and 0.07 in 2014 [65].

Moreover, CNN-based feature extractors were publicly released. This meant that the use of CNN-based features became available without requiring the knowledge or computing infrastructure for training a convolutional neural network from scratch. Among the first tools that were made publicly available we can cite *Overfeat* [14], followed by *Caffe* [81]. These extractors provide pre-trained weights files and facilitate the extraction of features for new image collections. The outputs of their final layer are semantic image representations but they are limited to the 1,000 ILSVRC concepts, due to computational complexity of the algorithm.

Recently, the authors of [318] and [319] exploit CNN to build mid-level features and report impressive results on various image classification datasets. For instance, performance improvements are achieved by [318] on PascalVOC 2007 dataset with a MAP of 0.777 compared to 0.705 for previous methods [320]. Recently, the use of CNNs has further increased the MAP score on PascalVOC 2007 to 0.824 [13], 0.852 [321], 0.897 [65], 0.906 [322], and 0.925 [323]. The focus here is not on building new CNN representations but rather on exploiting them as basic features in order to build powerful semantic representations from very large Web datasets. Although it has been proven that training a network using data for a specific domain rather than using one trained on ImageNet benefits feature transfer in the confined setting of that specific domain [324], in our work we use image representation stemming from three Imageet based models. This choice is motivated by the fact that our work deals with a broad spectrum of concepts. We look to find a common representation suited for a diverse large set of concepts rather than to tailor task specific descriptors. Our hypothesis is that CNN complexity limitations can be compensated by an appropriate choice of a conceptual support from au automatic processing of large-scale Web datasets.

We do not go here into the technical details on training CNNs. However, we offer a general image on how feature transfer is performed. Regardless on the network's architecture (e.g. the number of convolutional of fully-connected layers, the size of of the sliding window

used for the convolution operation, the image transformations or the use of regularization techniques, such as dropout), the global idea of this work is that the internal layers of the CNN can act as a generic extractor of mid-level image representations. The network can be pre-trained on one dataset (the source task, most common ImageNet) and then re-used on other target tasks, as illustrated in Figure 3.1.

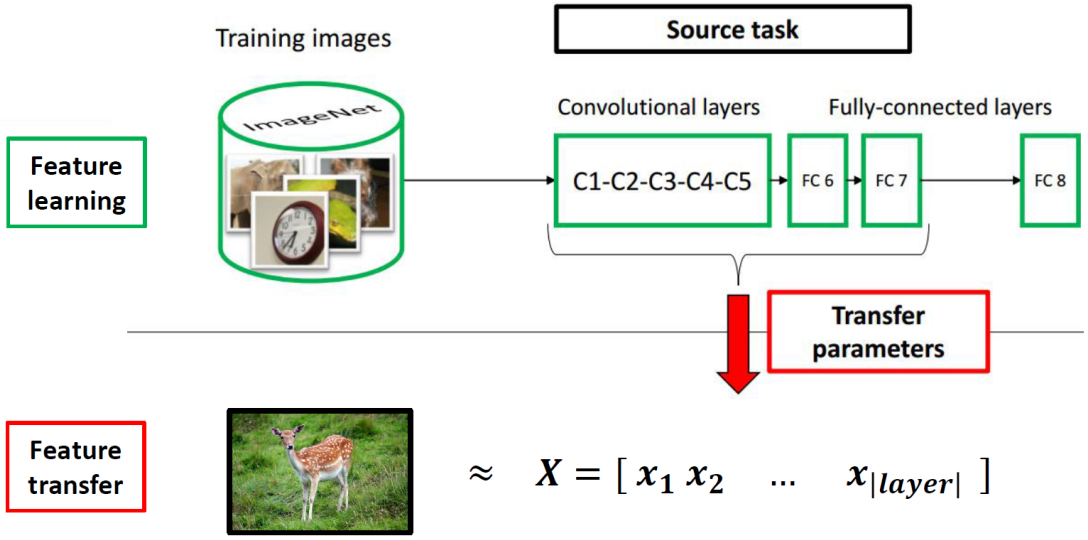


FIGURE 3.1: Transferring parameters of a CNN. Adaptation of the framework depicted in [31]. First, the network is trained on the source task (ImageNet classification, top row) with a large amount of available labeled images. Pre-trained parameters of the internal layers of the network are then used as descriptors for images in different tasks.

The network presented in this figure 3.1 a simplified view of the network introduced by Krizhevsky et al. in [63]. The size of the parameter vectors obtained after each stage is the following: *C1:253440, C2:186624, C3:64896, C4:64896, C5:43264, fc6:4096, fc7:4096, fc8:1000.* The last fully connected layer (i.e. *fc8*) offers predictions for the classes on which the network was trained upon. In order to get normalized prediction, a *softmax* function is applied to the output of this layer. In our example, where the network was trained on the ImageNet challenge data, the last layer will provide prediction for the 1000 classes used in the competition. In practice, the two fully connected layers are often used for feature transfer [318, 319], although features extracted from the first convolutional layers have been used for some tasks, such as texture detection.

3.2.2 Datasets preprocessing

In this subsection, we present the initial visual preprocessing step that focuses on the extraction of the image descriptors used for visual concept building. For the experiments described in the following sections, we compare two slightly different implementations of

the same convolutional neural network and also two network architectures implemented in the same framework.

The first framework that we used for extracting CNN features is Overfeat [14]. Two models are provided, a faster and and a slower, more accurate one. Each architecture is based in the model introduced by Krizhevsky et al. [63], with the faster network being more similar to the original. The slower model is more accurate than the fast one (14.18% classification error as opposed to 16.39% in the ImageNet Large Scale Visual Recognition Challenge 2013 (ILSVRC2013)), however it requires nearly twice as many connections. Although there is this difference in classification performance, in terms of feature transfer, there is little to no difference between the two networks and the extraction cost is unjustifiable, notably when applied at large scale. Equivalent performance was obtained in preliminary experiments not reported in more details here.

The second framework is Caffe [81]. It has recently developed into an open source project, which receives constant contributions from a supporting community. The framework is a BSD-licensed C++ library with Python and MATLAB bindings for training and deploying general purpose convolutional neural networks and other deep models efficiently on commodity architectures. Caffe allows CUDA GPU computation and can process over 40 million images a day on a single K40 or Titan GPU. By separating model representation from actual implementation, Caffe allows experimentation and seamless switching among platforms for ease of development and deployment from prototyping machines to cloud environments. This architectural choice represents one of the main advantages of this framework compared to Overfeat. It allows researches to publish newly proposed network architectures and even trained network parameters independent of the implementation. For example, among others, architectures or trained models are made available for the networks proposed by: Chatfield et al. [13], Long et al. [78], Zhou et al. [324], Lin et al. [325]. These were used for several tasks, such as: multi-concept image classification, object detection, scene classification or image retrieval. This increases the reproducibility of research and easy experimentation with recent advancements.

Next, we detail three configurations that we have tested for the experiments described in this Chapter and that we also use as baseline image representations in Chapters 4 and 5

- **Overfeat** The default configuration, i.e. layer 19 of the small, faster network provided by Overfeat, is used for representing the datasets and for experiments. In Overfeat, intermediate layers (e.g pooling or rectified linear unit layers) are counted. This means that layer 19 from the Overfeat implementation corresponds to the output of the last fully connected layer (i.e. $fc7$) from Figure 3.1. As

presented in the previous section, this leads to all images being represented by a vector of 4096 dimensions that is further normalized using L2.

- **Caffe** Through this Thesis, we will simply refer as *Caffe* when using as image representation the standard model provided by Caffe (*i.e.* the original model of Krizhevsky et al. [63]). Similar to Overfeat, we extract the weights of the *fc7* layer, followed by a L2 normalization.

- **VGG** We will refer by *VGG* the Caffe models of the networks described in [65]. These models are the improved versions of the models used by the VGG team in the ILSVRC-2014 competition [326] and are based on the observations made in [13]. Two models are made public: a 16-layer model (with 7.5% top-5 error on ILSVRC-2012-val and 7.4% top-5 error on ILSVRC-2012-test) and a 19-layer model (with 7.5% top-5 error on ILSVRC-2012-val and 7.3% top-5 error on ILSVRC-2012-test).

While for the experiments presented in this Chapter we rely solely on CNN features, it is worth mentioning that our initial work towards building large collections of visual concept classifiers was carried out using SIFT descriptors [47] aggregated into bags of visual words as low-level image representation. When comparing the cross-validation accuracy of individual models, the SIFT based models yielded considerably lower accuracy scores to those based in Overfeat features, thus confirming previously published results [79].

3.3 Visual concept learning

In this Section, we give an overview description of the framework that we propose for building visual concept classifiers. After the choice of an image representation, we are interested in an approach that focuses in speed and extendability, that also offers a high prediction accuracy. Considering the size of the problem we tackle, we use a set of binary classifiers to model visual concepts. In comparison with a multiclass classifier, this choice presents the advantages of (i) remaining computationally feasible for any number of classes and having lower constraints on the training dataset size and (ii) being easily *extendable* in the sense that adding (or removing) a given concept can be done independently from other models that have already been trained. For the sake of scalability, each concept is modeled with linear models, which are very fast to compute and exhibit good performance in practice[25, 64]. Hence, each individual model is learned from a set $\{(I_i, y_i)\}_{i=1...N}$ of training images and their corresponding binary label ($y_i \in \{-1, +1\}$). Models built with Overfeat features are learned with L2-regularized logistic regression, which solves the following unconstrained optimization problem:

$$W^c = \arg\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} log(1 + e^{-y_i\mathbf{w}^T x_i}) \tag{3.1}$$

Where $x_i \in \mathbb{R}^{S_f}$ is an image low-level feature reflecting the visual content of image I_i. This feature is later augmented with a last dimension fixed to 1 to take into account the model bias: $\mathbf{f}_i^T \leftarrow [x_i^T\ 1] \in \mathbb{R}^{(S_f+1)}$. In practice, (3.1) is solved in the primal using a trust region Newton method, relying on the liblinear implementation[327]. After a series of preliminary experiments, for the models built upon Caffe and VGG descriptors, we chose a L2-regularized L2-loss support vector classifier (*i.e.* the second solver from [327]). In order to normalize the prediction scores, we apply a *softmax* function.

The authors of Fernández-Delgado et al. [328] show that, when averaging the performance over a large number classification tasks, classifiers stepping from the random forest paradigm provide better classification performance than support vector machines classifiers. While this result is useful when choosing a classifier for a novel task, it is known that tree based ensemble models are considerably slower than linear models. Besides the gain in speed, linear classifiers have less parameters and are less sensitive to the choice of these parameters. Given that in this Thesis the emphasis is put on capacity of processing a large volume of image data, using a linear classifier is better suite for our needs. Also, our primary goal in not to fine-tune the individual performance of a classifier but to obtain classifiers that are useful in different tasks, such as semantic image description and user image tagging credibility estimation.

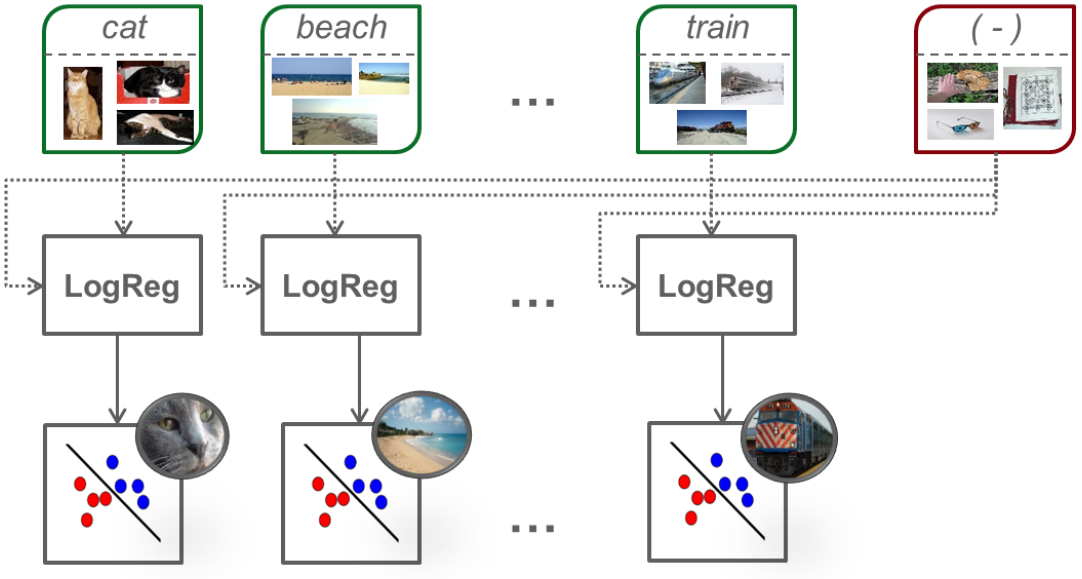


FIGURE 3.2: Individual visual concept models training framework.

In Figure 3.2, we show a graphical representation of the proposed framework for obtaining a set of visual concept classifiers. This framework can be used with positive training data coming from either manually labeled datasets (Section 3.4) or Web data (Section 3.5). An important choice when building a binary classification model is the choice of the negative class. One commonly used method is the on-versus-rest approach. This is mostly used in multiclass classification problems that deal with a small number of classes, such as PASCAL VOC [329]. In that case, for a fixed class, all the images coming from the other classes are taken as negatives in the classification process. Another approach, such as the one proposed by Li and Snoek [330] is to use a negative selection algorithm that chooses a different set of negative examples from a large set of images for each class. While this method can bring an increase in classification accuracy, the computational cost in non-negligible. Taking into account these observations, we opted for a **single large negative class** that will be used for training all concept models. Having this fixed, remains the question of the number of negatives that should be used for classification. While when performing binary classification, a balanced number of positive and negative training examples is proffered, we evaluate classifiers build using negative sets of different sizes. These experiments are detailed in Section 3.4.2. Finally, the weighs learned by the models, together with the bias term, are stored for further use.

3.4 Use of available annotated image resources

In this Section, we apply our visual concept modeling framework using manually labeled data. Our aim is to obtain a large set of concept models. ImageNet [12] is larger in scale and diversity than the other image classification datasets (*e.g.* [8, 329, 331]) and is thus fitted for use here. Next, we will describe this dataset, its structure, previous use scenarios and we will offer an exploratory analysis of the visual models obtained from ImageNet data.

3.4.1 ImageNet: A Large-Scale Image Database

ImageNet [12] is a visual resource which was built on top of the hierarchical structure of WordNet [332]. It contains manually labeled examples for 21,841 which contain a total of 14,197,122 images. The candidate images were collected from the Internet by querying several image search engines. For each synset, the queries represent the set of WordNet synonyms. In order to improve the accuracy of the dataset, the authors relied on crowdsourcing to verify each candidate image collected in the previous step for a given synset. The Amazon Mechanical Turk (AMT) was used to recruit workers and perform the annotation of images. ImageNet gained recognition mainly through the The

ImageNet Large Scale Visual Recognition Challenge [326]. This challenge is a benchmark in object category classification and detection on a thousand of object categories and over a million images. The challenge has been run annually from 2010 to 2015 (current year of writing) and has gained popularity leading to the participation of than fifty institutions in 2014. The publicly released dataset contains a set of manually annotated training images for 1000 concepts.

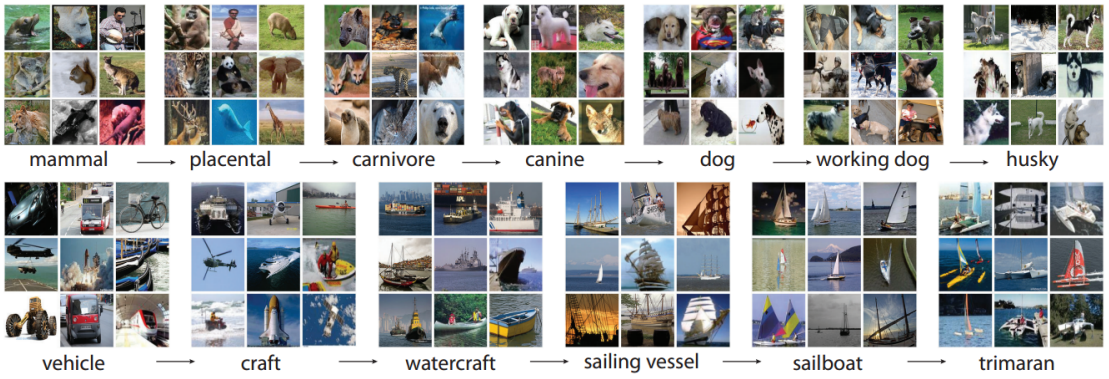


FIGURE 3.3: A snapshot of two root-to-leaf branches of ImageNet taken from [12]: the top row is from the mammal subtree; the bottom row is from the vehicle subtree.

In Figure 3.3, we take the example provided by [12] to illustrate the hierarchical organization of ImageNet in correspondence with Wordnet. Knowing that not all Wordnet concepts are represented in ImageNet, the hierarchy of the latter resource has a smaller size. For instance, the top 2 levels of concepts from Wordnet are not represented in ImageNet, namely the root concept (*i.e.* *entity*) and the second level concepts (*i.e.* *Physical_Object*, *Nonessential*). The highest level concepts from ImageNet correspond to those found on the third level of the Wordnet hierarchy, with the exception of the *Misc* class, which can contain concepts from different levels of the Wordnet hierarchy. In total, including the *Misc* class, there are 9 concepts at the top level in ImageNet. In Figure 3.3, the *mammal* and *vehicle* concepts are found on the third level, while *husky* and *trimaran* are situated in the ninth level. We can also see here that ImageNet provides images for all the intermediate concepts. This allows us to obtain visual representations across different granularities.

3.4.2 ImageNet based visual concept classifiers

In this Subsection, we detail the selection of ImageNet concepts for which we build classifiers and we study the link between the visual coherence of the concepts and their position in the Wordnet hierarchy.

TABLE 3.1: Examples of ImageNet concepts with high cross-validation scores (left column) and concepts with low cross-validation scores (right column).

Concepts with high CV scores	**Concepts with low CV scores**
bookcase bluetick snow_ leopard	*successor plainsman pessimist*
bison hatchback Model_ T	*Penobscot kleptomaniac experimenter*
white_ stork police_ van	*middle-aged_ man seeker stranger*
Old_ English_ sheepdog	*color-blind _person field_ pea*
black_ stork web_ site	*humanity part-timer*
blue_ point_ Siamese	*nondescript greenishness man*
Persian_ cat drake leopard	*monster water_ locust wonderer*
geyser scaup_ duck	*neutral scientist junior bankrup*
manhole_ cover dogsled	*Andorran witness failure*
subcompact_ car	

Our objective is to build a large set of visual classifiers from this resource and we do not constrain ourselves to using only the 1000 used in the ImageNet challenges. From ImageNet, we selected the **17,462** concepts which have at least 100 associated images and the resulting subset includes around **13 million** images. For the negative class, we take all of the images that come from concepts with less than 100 images. In order to ensure diversity for the negative instances, we then shuffle the images before indexing them. We investigate the visual coherence of obtained visual concepts by looking at the 5-fold cross-validation (CV) accuracy of the model trained using the concept's ImageNet images and a negative class comprising of the same number of negatives as are positives. For instance, if a concepts has k images in ImageNet, we extract the first k images from the large negative class and feed the positive and negative examples to the classifier. Although we build models for all of these concepts using all of the three image features detailed in Section 3.2.2, we perform the CV analysis with models built upon Overfeat features. Repeating the same experiment for 17,000 concepts with models built with other features would be time consuming and would not reveal novel insights, as we are not interested here to compare the features themselves.

We want to provide insight about visual coherence of concepts and therefore rank them in accordance to their respective scores. In Table 3.1, we show in the left column a list of ImageNet concept found among the first 1% in the ranked list (*i.e.* those with high CV scores) and in the right column, concept found among the last 1% concepts in the ranked list (*i.e.* concepts with low CV scores). We can observe here that, as expected, concepts with high CV scores are more specific concepts depicting, for example, types of cars (*e.g. police_ van, subcompact_ car*) or breeds of animals (*Persian_ cat, scaup_ duck*). On the other hand, concepts for which we obtain low CV accuracy scores cover general concepts. One trend that stands is the predominance of concepts visually depicting people, either a specific nationality (*e.g. Andorran*), a trait (*e.g. witness, neutral*) or a job (*e.g.*

scientist). In this category, we also find concepts focusing on visual information (*e.g. greenishness*) or abstract concepts without a clear visual representation (*e.g. failure*).

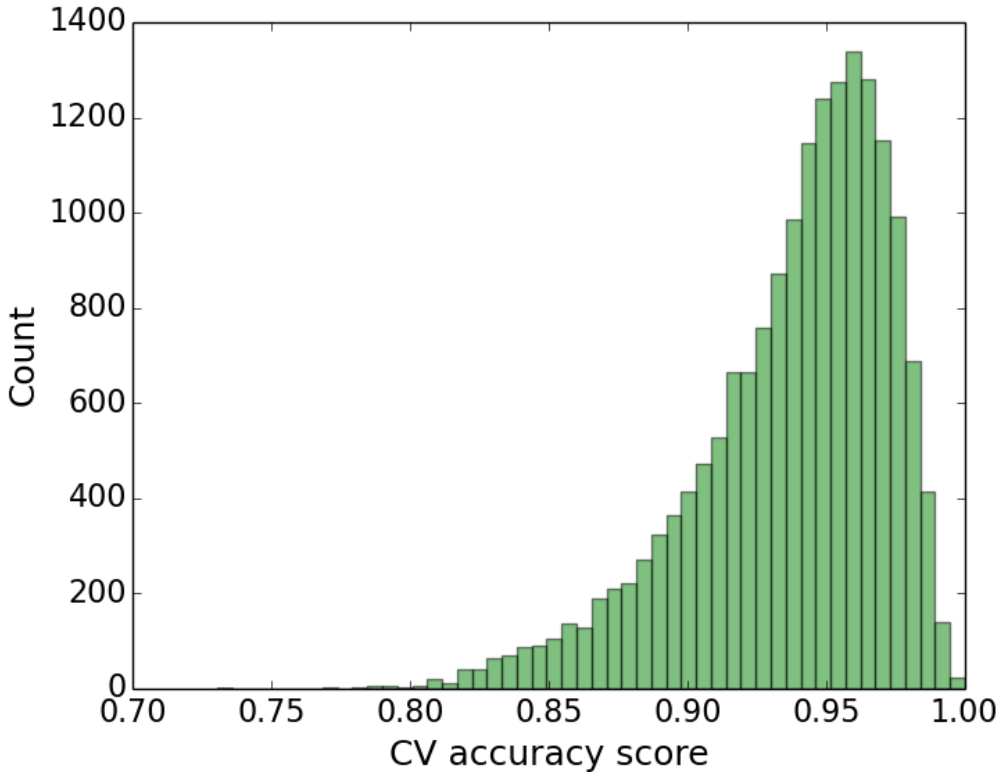


FIGURE 3.4: Distribution of cross-validation accuracy scores for visual models built from ImageNet concepts.

In Figure 3.4, we present a histogram of cross-validation accuracy scores for visual models built from ImageNet concepts. We can immediately observe that using a simple classification method, including a linear classifier, we obtain accurate concept detectors. We have a left tail distribution with most of the accuracy scores being over 0.9. The mean CV accuracy score is $\mu = 0.937$, with a standard deviation $\sigma = 0.038$. Besides the training images being manually labeled, another fact that can explain these high scores is that in the ImageNet images, the concept is clearly depicted. It is generally in the focus of the image, with little background or other obvious concepts.

Although there is not a big variance between the CV scores, we are interested in investigating what are the possible factors that influence the usefulness of manually validated image sets for visual concept modeling. A first observation is that abstract concepts are inherently visually diverse. We showed examples of such concepts in the second column of Table 3.1. A complementary hypothesis is that, except for visually diverse concepts, the position of a concept in the Wordnet hierarchy influences the quality of the visual model. In Figure 3.5, we show a box plot in which we showcase the distribution of CV

scores with respect to the depth in the Wordnet hierarchy. As mentioned in Section 3.4.1, the highest position of an ImageNet concept is 3 in the Wordnet hierarchy and it can go up to 18. As expected, we can see a lower variability in CV scores for higher level concepts (*i.e.* levels 3 - 7 in the hierarchy) and lower level ones (*i.e.* levels 14 - 18), while middle-leveled concepts display higher variability and a larger number of outliers. This is mainly due to the fact that there are fewer concepts found towards the two extremes of the hierarchy. In order to emphasize the correlation between the cross-validation scores and the position in the hierarchy, we have also plotted the mean values for the concepts found on the same level (*i.e.* the blue points inside the boxes). Looking at the mean values, we can observe an almost monotonous increase of accuracy scores when the position in the Wordnet hierarchy lowers up to level 14. Even if the CV value does not highly increases when passing from one level to another, this confirms our intuition that there is a relation between a concept's the position in the Wordnet hierarchy and its visual homogeneity.

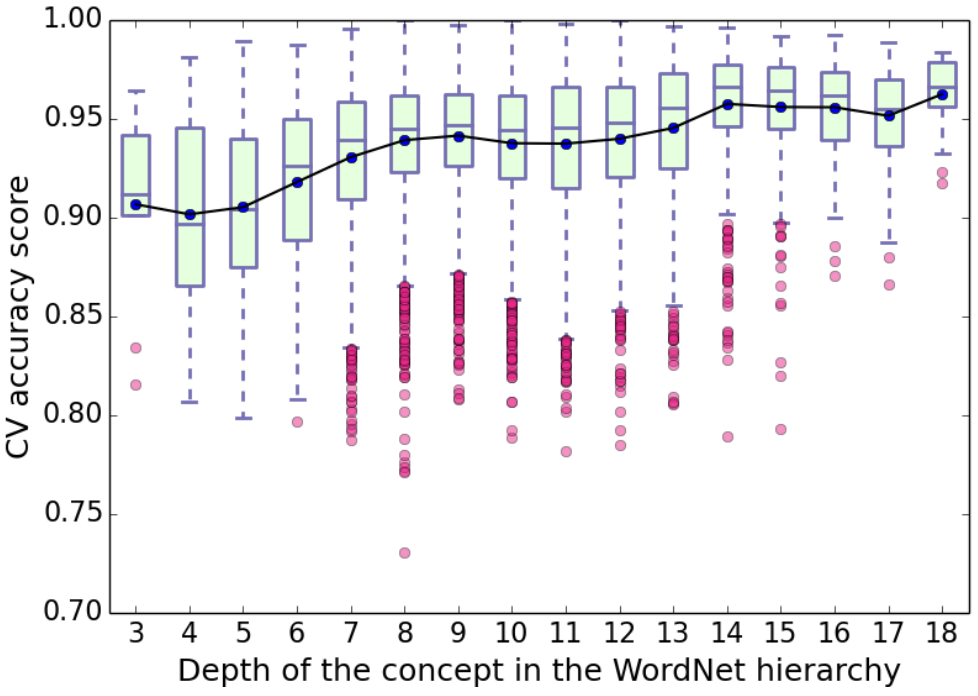


FIGURE 3.5: Correlation between the depth of concepts in the Wordnet hierarchy and the cross-validation accuracy score of the visual model built from them.

The results presented in Figure 3.5 complement the findings reported in Deselaers and Ferrari [333]. These authors investigate how semantic distances between categories defined on the WordNet hierarchy relate to visual distances in ImageNet. They measure the visual distance between two concepts as the average distance between the mean descriptor of the first one and all images in the second one. The semantic distance is measured

using the Jiang and Conrath distance [334] and results show that the visual distance continuously grows with the semantic distance. They conclude that visual similarity as measured by computer vision descriptors conveys semantic similarity, analog to what shown for human perception.

3.5 Dealing with noisy Web data

We are interested in this Section to exploit social intelligence for gathering a large collection of Web images that is exploited to build visual concept models. While most of the works that collect Web images for image processing tasks use image search engines, such as Google or Bing, our focus is directed towards Flickr groups. When using search engines, the origin of the image is not persevered and we rely on the search algorithms to return relevant examples. Also, we would need a fixed list of concepts to launch a text based query. Since images found in Flickr groups are gathered around users' interests, these groups provide a *natural* organization of concepts in an image sharing platform, without the intervention of Flickr's internal ranking mechanisms. Groups can either be curated by the group's creator, by several users or they can have a more permissive policy towards accepting contributions from users. The authors of Negoescu and Gatica-Perez [335] consider that Flickr groups offer viable new alternatives to organize and manage visual content. They are self-organized communities with common interests that are created spontaneously but not randomly: people participate in groups (e.g. by sharing pictures) for specific social reasons, and most groups are revolve around specific topics or themes of interest (e.g. an object, an event or a photographic style). Aggregating content and metadata for groups could offer insights into both large scale behavioral trends (e.g. photo sharing practices), and also provide robust representations (e.g. at the topic level) to characterize groups by their content [335].

We collected Flickr groups starting with an initial list of 100 million images from which we extracted the most frequently occurring groups. Then we downloaded group metadata for the most frequent 50,000 of them and retained the **38,500 groups** which include at least 300 images. Given that some images were withdrawn by users before crawling and that a part of the images appear in several groups, the initial dataset contains approximately **11 million** images. This selection is done in order to ensure that a reasonable amount of data is fed into the visual classifiers which are build from Flickr groups.

3.5.1 Flickr group modeling

Similar to the processing of the ImageNet images, we extract all of the three image features detailed in Subsection 3.2.2 for Flickr group images. To maintain comparability with ImageNet models, we use the same large class of ImageNet images as negatives (*i.e.* the negative class described in 3.4.2). Similarly, we train models using all three image features but we perform the CV analysis only with models built upon Overfeat features.

While the meaning of each concept is known for ImageNet, in the case of Flickr groups, a first challenge is finding the proper textual description of the group that best describes its visual content. A first possible choice is the group's title. After an investigation of a set of group titles, we noticed a high level of noise among titles (*i.e.* non alpha-numeric characters, different languages, subjective statements). There are also a lot of titles that have a narrative nature, making them impractical for a proper textual representation of the group. Another problem with choosing titles is the bias towards the initial choice of a group's author. This is notably encountered among less carefully curated groups, where the content may evolve in another direction than the one initially intended by its creator. For all of these reasons, we chose a data driven approach that is based on the predominant tags associated to the images found in the groups. In this way, we also capture the collective social intention behind tagging for the set of user that provided contributions to the group. Text pre-processing consists in extracting the most salient tags of each group. Groups are structured thematically but a single tag might not be sufficient to describe them. Tags are ranked by the number of unique users which annotate images of a group with them. This measure is chosen instead of tag frequency, which is sensitive to bulk uploads, in order to maximize the social relevance of tags. In this way, we eliminate the possibility that a single user would have a high influence over the tags selected for describing the group. After an initial examination, we empirically retain the top three tags as a textual representation of groups and write this representation as $FG_t = \{T_1, T_2, T_3\}$.

In Figure 3.6, we present a histogram of cross-validation accuracy scores for visual models built from 38,500 Flickr groups. For training these visual models, we used a balanced training set (*i.e.* the same number of negatives as there are positives for each concept). We can observe that, although we are using noisy Web images and a simple classification method, including a linear classifier, we obtain accurate concept detectors. The mean CV accuracy score is $\mu = 0.904$, with a standard deviation $\sigma = 0.051$. When comparing with the distribution of ImageNet based models CV scores (Figure 3.4), we note that the distribution of Flickr groups is more disperse. One explanation for this difference comes from the fact that we have 38,500 groups models versus approximately 17,000 ImageNet models. What is of most interest to us is that we still have a large number of groups

models with scores over 0.9. Seeing that the average CV score of Flickr groups is only 3% lower than the one obtained for ImageNet, we already have an important clue on the usefulness of groups as positive training images for building visual concept classifiers.

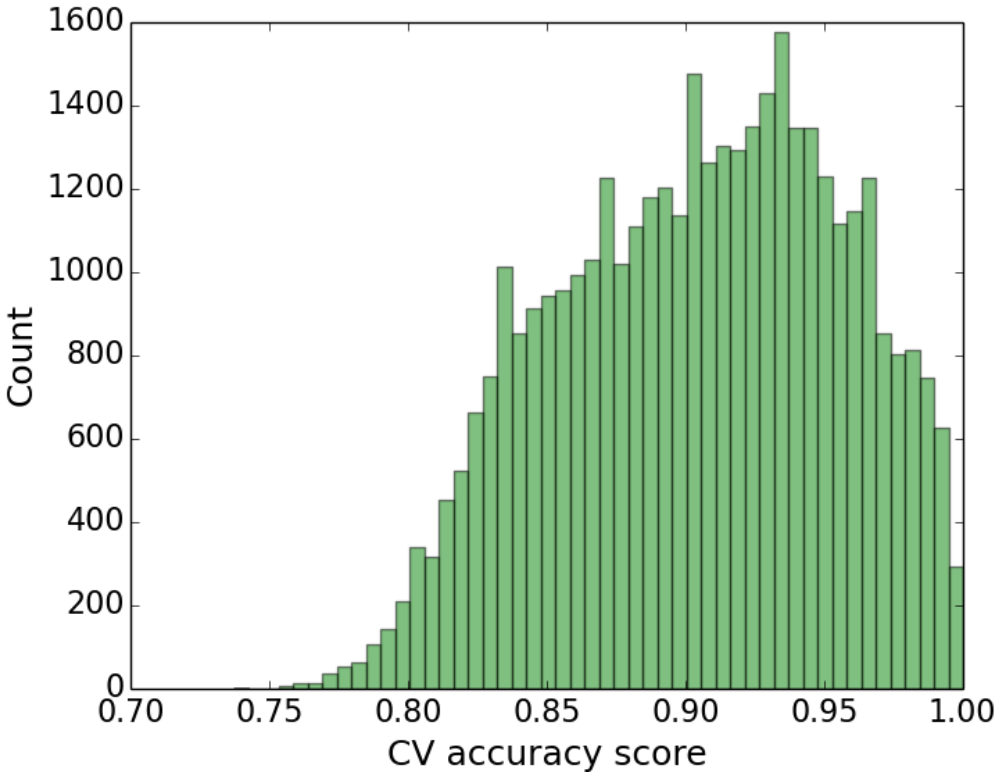


FIGURE 3.6: Distribution of cross-validation accuracy scores for visual models built from Flickr groups.

Flickr groups may form around specific concepts (brands of cars, animals etc.), abstract concepts (beauty, frightening imagery) or they may gather images taken with a specific brand of camera or camera setting (black and white, light setting). In Figure 3.7, we provide an example of two types of groups. The upper part of the image contains samples taken from a group formed around the concept *truck*, which has a clear visual representation. The lower part of the image contains samples taken from a group formed around the concept *dreamy*. Our goal is to obtain a large set of visual concept models that, while diverse, covers concepts with a clear visual representation. We are therefore interested in selecting groups for which we can find a coherent view for their images. Next, we describe the method that we propose in order to perform this filtering.

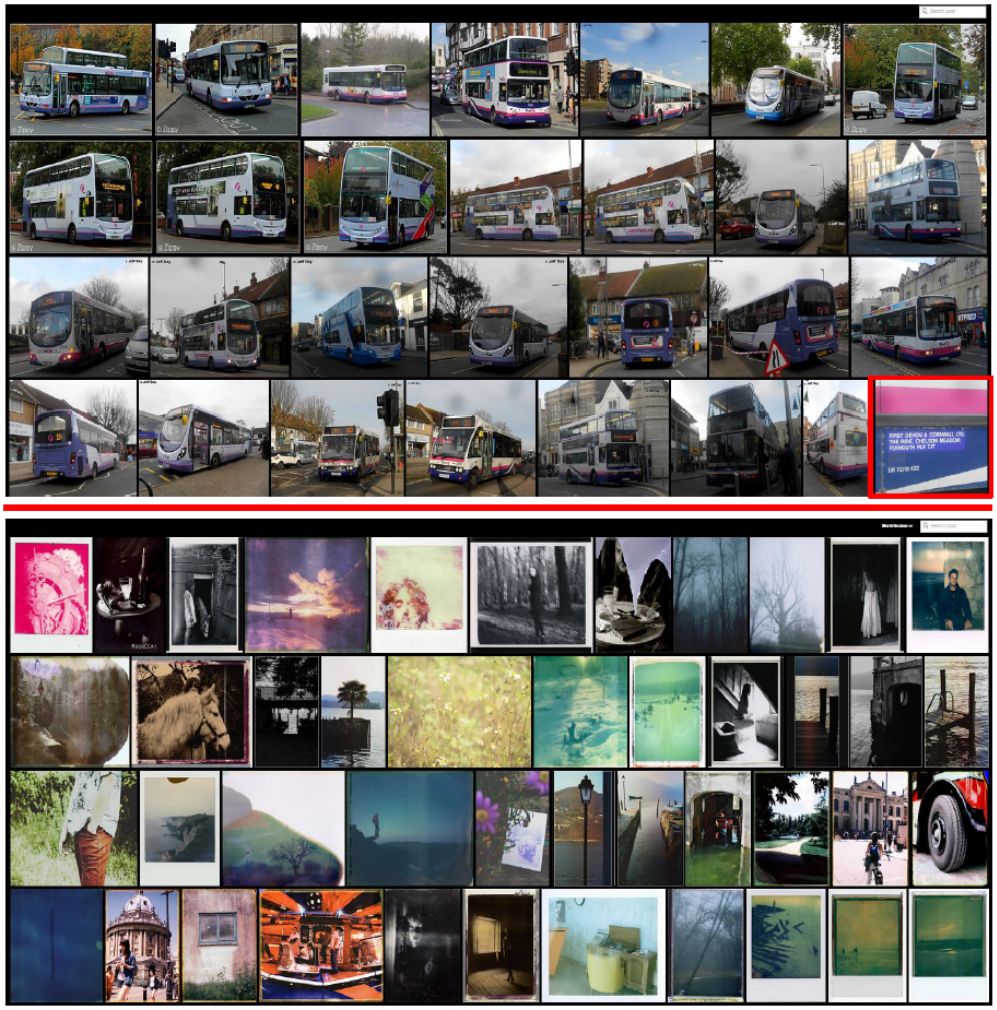


FIGURE 3.7: Example of a visually coherent group (upper image) and a visually incoherent one (lower image). The upper image contains samples taken from a group formed around the concept *truck*, which has a clear visual representation. The lower image contains samples taken from a group formed around the concept *dreamy*.

We investigate how much visual variability there is among the images of a group. We measure the visual coherence of a group by the 5-fold cross-validation score reported by the model trained on that group, as described in Section 3.3. Figure 3.8 confirms that the ranking based on the cross-validation score manages to separate groups depicting concepts with a clear visual representation from those focusing on photography or generic concepts. Only the first 30,000 groups ranked by cross-validation are retained for the experiments described in Chapters 4, 5 and 6. Most other groups are not conceptually oriented and have little added value for visual concept building. Taking the average CV scores of the first 30,000 groups, we notice a 2% increase, when compared to the average of the full set of groups models. The mean CV accuracy score for the first 30,000 groups is $\mu = 0.924$, with a standard deviation $\sigma = 0.036$. This also entails that the mean CV score for this set of groups is only 1% lower than that for ImageNet models.

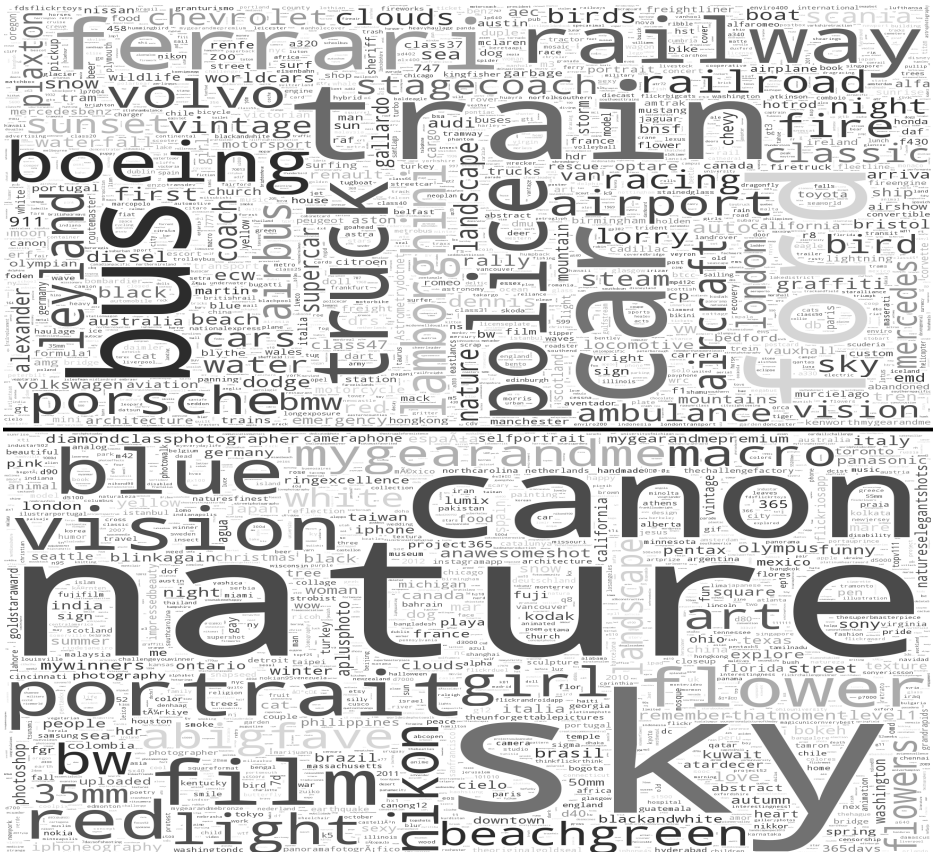


FIGURE 3.8: Word clouds of the most frequent tags found in the first 10% groups (upper word cloud) and the last 10% groups (lower word cloud) in a ranking induced by the cross-validation accuracy score.

The performance of semantic descriptors obtained from the aggregation of visual models from ImageNet and Flickr groups are compared throughout the Chapter 4. For a better understanding of the results, we are interested in the semantic overlap between ImageNet concepts and groups and the particularities of each data source. We consider that an ImageNet concept and a group match if at least one term describing the concept has an exact match in the FG_t representation of the group. From the total of 17,462 ImageNet concept names, only 2,567 are found in groups. The concepts with the a lot of associated images in ImageNet that do not appear in groups include species of animals *African elephant, eastern gray squirrel*) or technical equipment (*computer keyboard, microphone*). When first looking at Flickr groups, we find 28,243 groups that have at least one tag matching an ImageNet concept. Among the first groups ranked by the number of contributing users that do not have an ImageNet correspondent, we notice a high frequency of geographical locations (*{paris, france, eiffel}, {croatia, sea, dubrovnik}*) and car brands (*{bugatti, veyron, supercar}, {lamborghini, gallardo, murcielago}*). This finding suggests that Flickr groups cover mainly the general concepts from ImageNet but also contain a high variety of specific concepts. The comparison of ImageNet and Flickr

groups confirms that most common concepts are covered by the two resources. However, when it comes to specific concepts, ImageNet covers specialized taxonomic concepts, while Flick groups mostly cover named entities which match users' interest and are not modeled in ImageNet. The main advantage over ImageNet comes from the nature of the concepts found in groups that are formed through social consensus, as opposite to ImageNet, where the specific concepts come from the leafs of the WordNet hierarchy and may be less frequently represented in the images shared through online platforms.

We also examine the semantic coverage of Flickr groups. From the complete set of 38,500 groups, we find 13,488 unique tags among the first three tags of each group. This number and the high proportion of tags found in ImageNet support the idea that popular tags in a group are likely to correspond to higher level concepts (e.g. concepts that are found on the upper levels of the WordNet hierarchy). Subconcepts can be identified by inspecting the accompanying tags. For example, if we look for groups containing the tag *cat*, we encounter groups built around specific species of cats: *{cat, britishshorthair, british}*, *{siamese, cat, lilac}* and *{cat, black, blackcat}*. This indicates a semantic hierarchical organization of groups, a finding reported in [336].

3.5.2 Group image reranking

A part of the images associated to Flickr groups are irrelevant and direct learning of visual models with all group images is probably sub-optimal. As it can be seen in the upper part of Figure 3.7, although we present there a visually coherent group, there is still some amount of noise (*i.e.* the image emphasized by a red square). We introduce image reranking techniques in order to automatically reduce the amount of noise present in groups. Existing approaches exploit tags [337] or rely both on tagging and/or visual content [338] but we are mostly interested in the visual aspect. We test two classical methods and also introduce one which gives an important role to social cues. Flickr images come with a wealth of metadata and, with [339], we hypothesize that cues such as the identity of the uploader can be exploited in order to improve image reranking algorithms. Focus is put on scalability in order to be able to process Flickr groups efficiently. All methods use the Overfeat representation of group images described in 3.2.2 and we implemented:

- avg_{sim} - this baseline method computes I_{avg}, the average Overfeat representation of each group, and ranks the images of the group by considering $sim(I_i, I_{avg}) = \frac{1}{||I_i - I_{avg}||^2}$, the inverse of the L2 distance from the average representation. The intuition supporting this method is that the similarity with group average is a good indicator of image relevance.

- *kNN* - classical method which compares group images with a set of diversified negatives in order to favor images which are best linked to other images of the same group. We keep computation cost low by choosing as many negative as there are images in the target group. The negative set size is chosen to be similar to that of the group in order to minimize the separation margin between the two classes. Negative examples are selected from groups which are visually similar with the target group. Visual similarity between groups is computed with $sim(I_{avg}, J_{avg})$, where I_{avg} and J_{avg} are the average representations of the groups to be compared. However, since several groups can illustrate the same theme as the target group and choosing their images as negatives would not make sense. Consequently, we exclude all similar groups whose text representation FG_t has at least one common element with that of the target group. Then we build a negative set by uniformly sampling the most similar 100 remaining groups. The reranking score of each target image is given the position of the 10^{th} group image in the list of similar images which includes both positives and negatives. The higher this position is, the better the image rank will be. This reranking approach is motivated by the assumption that relevant images are more similar to other images of the group than to images from other groups.

- *skNN* - is a "social" version of *kNN* in which all images which come from the same user as the target image are excluded from the list of similar images. Here we assume that an image is more likely to be relevant if it is visually similar to images uploaded by other users. *skNN* is more robust to bulk upload behaviors than the simple *kNN* algorithm.

Training images are sorted according to one of the methods described above. Then only a part of reranked list, later referred as the percentage of the whole list size *cut*, is retained for group modeling. The impact of the proposed reranking methods is experimentally explored in Section 3.6.2.

3.6 Experiments

In this Section, we first evaluate the impact of the image representation used for training visual models and the influence of the number of negatives. Then, we evaluate the role of the reranking methods that were introduced in the previous Section.

3.6.1 Choosing the initial image descriptor and the number of negative examples

For the experiments described in this section, we randomly select 100 ImageNet concepts that have at least 200 images and use up to 100,000 negative examples taken from the large class of negatives presented in Section 3.4.2. From each ImageNet concept, we randomly select 100 images for test and use the rest of them as positive instances in training. Our main interest is the influence of the number of negatives that are included in the classification process. We perform tests with the following numbers of negative examples: the same number of negative examples as there are positive images for a concept (noted *pos*), at least 500 (we take the maximum value between 500 and the number of positive examples for a concept), at least 1000 (we take the maximum value between 1000 and the number of positive examples for a concept), 5000, 10,000, 25,000, 50,000, 100,000.

Given that Overfeat and Caffe features are based on similar implementation of the same convolutional neural network architecture (see 3.2.2), we chose to perform our tests with Caffe descriptors due to the higher flexibility offered by the extraction framework. Seeing that VGG features stem from a different network architecture, we compare Caffe with VGG throughout the experiments described in this Section. Logistic regression tuning is done via a grid search on the parameter C. The best average cross-validation accuracy classification scores are obtained with $C = 10$.

We are firstly interested to compare the performance of the proposed model configurations in differentiating between positive and negative examples. After we train the models, for each concept, we get prediction probability scores for the 100 positive test samples and a fixed set of 5000 randomly selected negatives. We ensure that the negative images used for test are different from the negatives used in training. Finally, we rank the 5100 images based on the prediction score. For the result presented in Figure 3.10, we use mean average precision at different cutting points ($MAP@k$).

We first observe that the only significant leap in performance is obtained when passing from the first 100 ranked images to the first 500. There are small differences between MAP@500, MAP@1000 and MAP@5000. This is expected, as we test with only 100 relevant images and we already obtain high MAP scores even at the 100 cutoff.

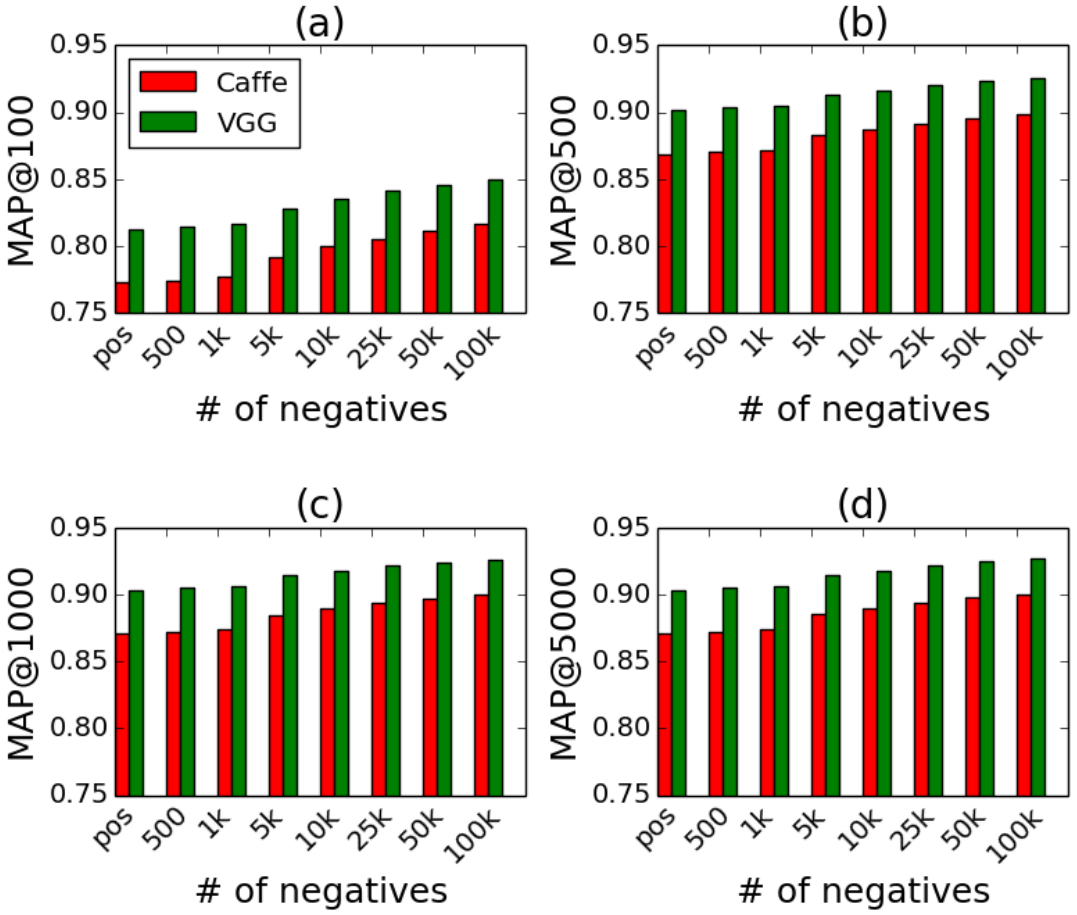


FIGURE 3.9: MAP scores of models trained using Cafee and VGG features as image representation and different number of negatives. We present results for MAP@100 in plot *(a)*, for MAP@500 in plot *(b)*, for MAP@1000 in plot *(c)* and for MAP@5000 in plot *(d)*. *pos* referrers to the case when there were taken as many negatives as there are positives for each concept. When the concept has more than 1000 positive examples, for the *500* and *1000* labels, we take the as many negatives as there are positives.

We can draw other two important conclusions from this experiment that serve as guidelines for large scale visual concept modeling:

- **VGG over Caffe**: When comparing the CNN features extracted from the two network configurations, the VGG features outperform the Caffe ones over all configurations and evaluation metrics. On average, we get a 2% increase of MAP@k scores.

- **Importance of the number of negatives**: We can observer a steady increase for the MAP@K score, for all K's in {100, 500, 1000, 5000} when the number of negatives used in training increases. This gain is more noticeable when looking at subplot *(a)* and for Caffe features. In this configuration, an increase of

approximately 4% of the MAP@100 score is obtained when using 100,000 negatives instead for training instead of choosing the same number of positives. While we obtain better classifiers with a higher number of negatives, there is a small increase in MAP@k when passing from 25,000 negatives to 50,000 or 100,000 in this preliminary experiment. When passing to large scale image retrieval with image descriptors obtained from the output of binary classifiers (Chapter 4), using a larger number of negatives plays a higher role in increasing retrieval performance.

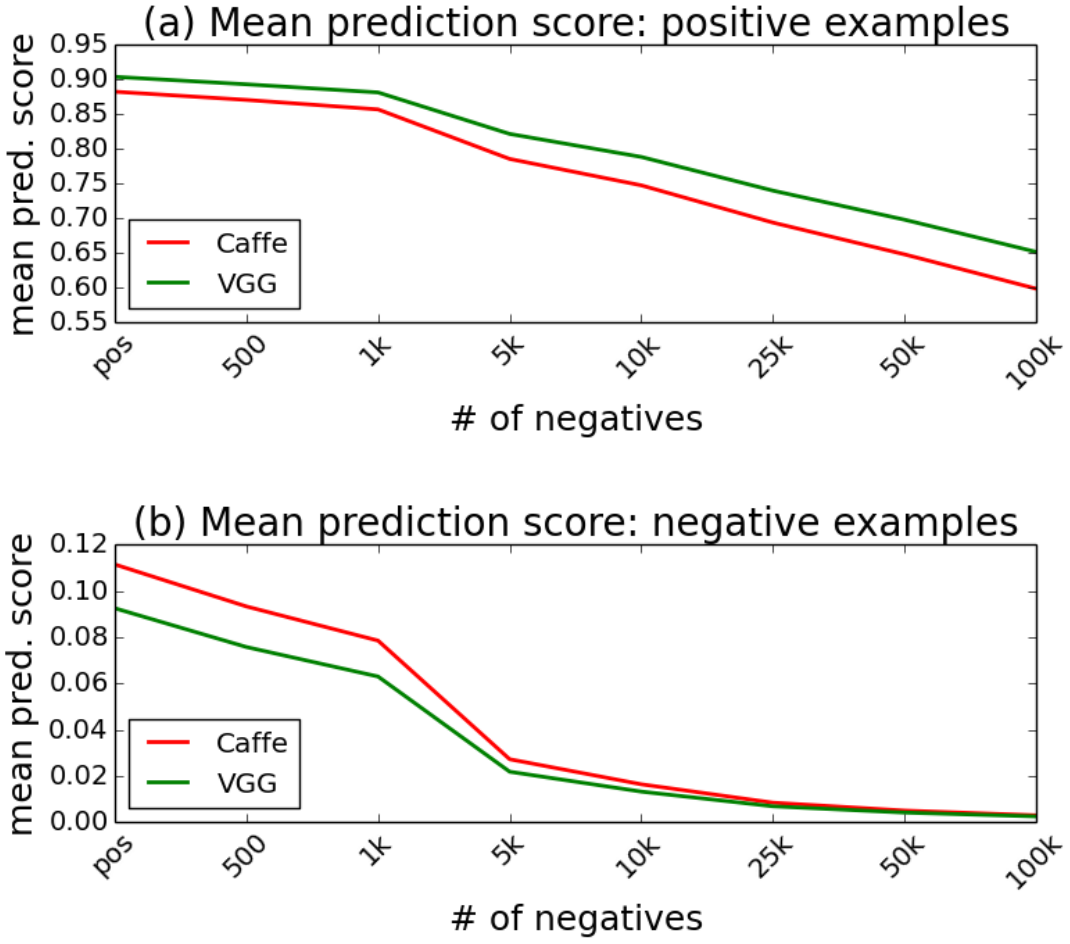


FIGURE 3.10: Mean prediction scores of models trained using Cafee and VGG features as image representation and different number of negatives. Mean prediction scores for 100 positive examples (plot *(a)*) and Mean prediction scores for 5000 negative examples (plot *(b)*)

In the following Chapters, we use the prediction scores of the individual classifiers, instead of the class prediction (positive or negative). For the logistic regression classifier, the standard cutoff point for class prediction is 0.5 (i.e. a prediction score over 0.5, will

classify the example as a positive and as a negative, otherwise). In Figure 3.10, we investigate the mean prediction scores for positive and negative test examples when using a different number of negative instances for training. We also compare the VGG and Caffe features. As expected, when having an imbalanced training set by adding more positives, we notice a drop in the prediction scores for positive test samples (subplot *(a)*). VGG features outperform the Caffe ones and the difference between the two descriptors becomes clearer with the increase in negatives. For 100,000 negatives, when using VGG features, the models have a prediction probability for the positive test samples ≈5% higher than models based on Caffe features. The same pattern is observed for the prediction scores for the 5000 diversified test negative set (subplot *(b)*). VGG based models give lower probabilities for the negative test images. The difference between the two features becomes negligible when using at least 5,000 negative training instances. It is important to notice here that, for both image descriptors, the prediction scores for the negative test images converge towards 0 with the increase of negative training images.

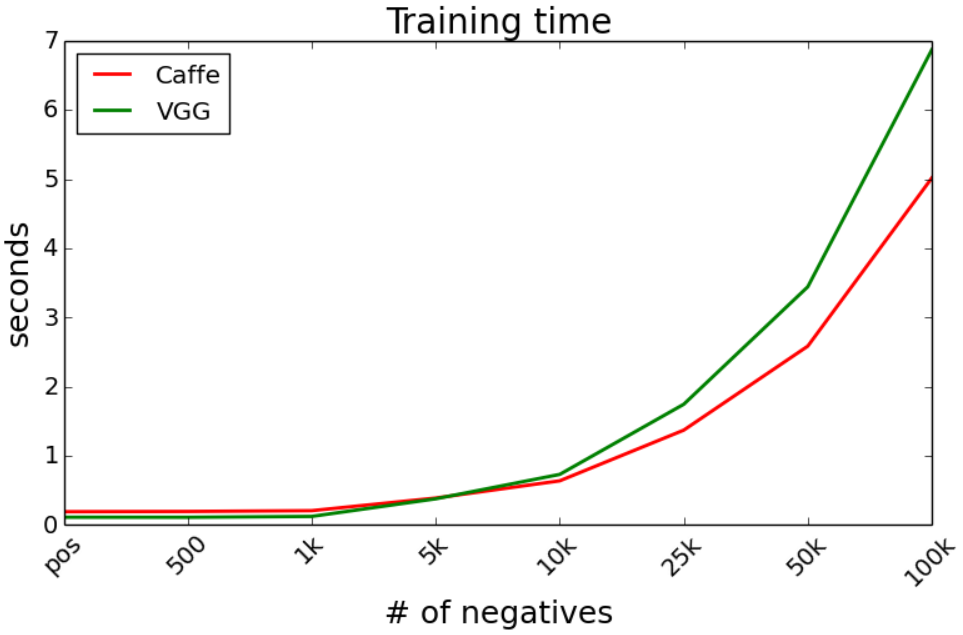


FIGURE 3.11: Mean training time (in seconds) for models trained using Cafee and VGG features as image representation and different number of negatives

In Figure 3.11, we compare the average training times for models trained using Cafee and VGG features as image representation and different number of negatives. This experiment is performed on a machine with an Inter Core i7-3630QM @2.4GHz CPU. For instance on a single core, we can train 17,000 ImageNet VGG based visual concept models in ≈1 hour when using 5,000 negatives and ≈33 hours with 100,000 negatives. We note here that the training times reported into this section refer solely for the model training process and do not take into account the time needed for loading the training data into

memory. In practice, when training all the concept models (i.e. 17,000 ImageNet and 30,000 Flickr groups), we bypass the issue of frequent disk access by working in a cluster environment with 256GB RAM per node.

3.6.2 The influence of Flickr groups image rereanking

The purpose of this preliminary experiment are to tune individual model learning algorithm and to evaluate results obtained with the different image reranking methods introduced in Section 3.5.2. The validation dataset used here is created by matching Flickr groups, used for training, with ImageNet concepts whose images are used for testing. We first pre-select a list of groups which have their first tag of their textual representation FG_t, present in ImageNet. For instance, we match the Flickr group 1000405@N24 (top tags *memorial, war, warmemorial*) with the ImageNet concept *memorial*, defined as *a structure erected to commemorate persons or events*. To ensure diversity, each tag is used only once. We manually validate the alignment between groups and ImageNet concepts to obtain a final list of 367 pairs. Training is done using Flickr group images as positive examples and a diversified negative set extracted from Flickr. Tests are performed with the images of the corresponding ImageNet concept as positives and a fixed list of over 4000 images of other concepts as negatives.

TABLE 3.2: P@100 results for different reranking methods introduced in Section 3.5.2 and different cut-off percentages (*cut*) for the selection of reranked images. The baseline corresponds to a no cut-off, i.e. the rightmost column.

	cut[%]			
	70	80	90	100
avg_{sim}	0.915	0.917	0.92	0.917
kNN	0.918	0.92	0.92	0.917
skNN	**0.922**	**0.921**	**0.922**	0.917

The reranking methods are compared by using them in an image retrieval scenario. For instance, assuming that *palm tree* is part of the 367 pairs, the purpose is to use the models trained with Flickr groups in order to classify ImageNet test images ImageNet and negatives. Using classification scores, we produce a ranking and assume that the best reranking method is the one which places the most *palm tree* among the top images. For each of the 367 Flickr group - ImageNet concept pairs, image classification scores are used to rank the test set. We classify all the images associated to the 367 ImageNet concepts and the negatives using the models trained on Flickr groups, with and without reranking. Then we rank positive and negative test images The precision at 100 (P@100), i.e. the number of positives among the first 100 results, is a used for assessment. We

choose this cut-off since the tested ImageNet concepts have at least 100 associated images. This measure accounts for the capacity of the reranking methods to favor positive test examples over negatives and, indirectly, for the quality of the reranked training set. Also, compared to classical cross validation, all reranking cut-offs are evaluated using the same test set and results are easier to compare.

The results obtained with the three reranking methods introduced in Section 3.5.2 and different cut-off points are presented in Table 3.2. While the P@100 differences with the baseline are small, some improvement is obtained with all reranking methods. More interestingly, the use of *skNN* provides slightly better results compared to *kNN*. This finding indicates that the use of social cues for reranking is beneficial for reranking performance. Following the results presented in Table 3.2, we will use *skNN* at different cut-off points for image retrieval experiments presented in Chapter 4.

3.7 Conclusion

The direct use of Web corpora for image mining, as proposed in [22] or [23], yields lower performance compared to manually curated datasets. However, we showed in this Chapter that with an appropriate choice of the initial collection and with the introduction of efficient image reranking techniques, the results obtained with the automatically built resource can rival with those of the manual resource. A good coverage of the conceptual space is obtained with an appropriate choice of the initial Web dataset. We explored the use of Flickr groups, but the pipeline presented here is easily applicable to larger datasets. The only potential constraints are the availability of data and the processing power needed to build individual models. We also investigated in this Chapter the choice of image descriptors and the number of negatives examples used for training the models. The experimental results serve as guidelines for the semantic features framework introduced in Chapter 4.

Chapter 4

Efficient CBIR with semantic descriptors

The main contribution of this chapter is an approach to design semantic image features, based on an array of individual concept classifiers built on top of automatically processed large-scale image collections. In a comprehensive experimental section, we show that the proposed descriptor improves the retrieval performance on three well known image collections (ImageCLEF Wikipedia Retrieval 2010, MIRFLICKR, NUS-WIDE), when compared to some widely used image features. We notably show that reducing the size of the semantic descriptors through sparsification not only increases retrieval performance but also accelerates the retrieval process through the representation of image collection with an inverted index.

4.1 Motivation

Multimedia data make for a large part of the content shared on-line and retrieval methods which combine effectiveness and scalability are needed to access these data under real time constraints. With the success of photo sharing platforms such as Flickr or Instagram, visual content has gained increasing importance on the Web. Responding to different practices, there are two main ways to access multimedia collections, namely through text and image queries. Text-based image retrieval, which is still predominant, is a subtype of text search in which results are returned by modeling textual information associated to images. It has been shown that the motivation of users has a major influence on the

annotations available on photo sharing sites [15]. Low motivation often leads to a poor or partial correspondence between the actual visual content and the associated textual annotations. Equally important, text search only provides access via matching with manual annotations, which are scarce or often missing. As a result, significant subsets of images are impossible to explore. Content based image retrieval (CBIR) is an alternative which is based on a low-level representation of images and requires no or little manual intervention. CBIR is already available in major search engines, such as Google or Bing. However, it suffers from the lack of coincidence between semantic image description and low level representations used to compute similarities, known as semantic gap [340]. In spite of important progress realized in visual mining during the last years, large-scale image retrieval is still mainly text-based.

We hypothesize that existing image retrieval systems can be improved through an appropriate exploitation of visual knowledge derived from large-scale resources. We propose a method which aggregates initial low-level image representations, such as *Caffe* or *VGG* (see Chapter 3), into an array of individual concept classifiers in order to compute a semantic representation of images. Semantic image descriptors are high level image representations which exploit a background visual resource to model concepts and assign conceptual representations to image collections [22]. As such, they have the potential to bridge the gap between content and semantic description.

In spite of sustained efforts and important progress over the last decade [341, 342], a number of important challenges still have to be addressed before including automatic annotation in Web retrieval pipelines. First, a very high number of concepts should be modeled to provide support for a large variety of user queries. Second, the quality of automatic concept labels should be close to that of human annotations. In this Chapter, we propose an image representation pipeline which partially addresses the limitations cited above. Scalability is obtained through the use of noisy but comprehensive background Web resources, while efficiency is ensured through image representation with convolutional neural network (CNN) features. Our main contribution is the design of a semantic image features which is built on top of an automatically processed large-scale collection. Since the resource is collected from Web images, a key part of the work deals with methods that reduce the effect on noise present in Web collections. Experimental validation mainly addresses content based image retrieval but also image classification. Globally, we address the following open and recurring research questions:
Q1 (manual versus automatic resources) - Do semantic features built on top of automatically and manually built resources have similar performances?
Q2 (semantic coverage) - Is it possible to build image representations whose components efficiently convey semantic meaning and ensure a good coverage of the semantic space?

Q3 (large-scale retrieval) - How to build semantic representations which are both compact and accurate?
Q4 (domain transfer) - Can we learn semantic features with a given resource and then successfully exploit them to mine other datasets?

Q1 is the central question addressed here and our approach is validated only if performances obtained with the two types of resources are comparable. To our knowledge, the direct comparison of automatically and manually built large-scale resources was not properly addressed in literature. *Q2* relates to the use of semantic features as an alternative to low or intermediate level features, such as bags-of-visual-words, Fisher Kernels[25] or CNN features[14]. Unlike low-level feature vectors, semantic features directly convey humanly understandable information. Consequently, they would be a promising candidate for bridging the semantic gap if they would be both precise and comprehensive, two conditions which are not yet met. For instance, the performances of meta-classes [23], which exploit a large part of ImageNet, lag behind those of Fisher Kernels, whose performances are in turn lower than those of CNN features [63]. We hypothesize that Flickr groups are a good candidate to answer *Q2*, provided that they are properly selected and cleaned via image reranking. Regarding *Q3*, we note that textual documents can be searched efficiently because they are sparse and a similar property is desirable for semantic image features. Surprisingly, sparsification is not directly addressed in existing work, with the closest proxy being quantization done for more efficient signature storage [23]. *Q4* is a major topic in computer vision that relates to transferring knowledge gained from a dataset to other datasets. It is crucial from a practical point of view, since one can not always adapt his/her system to a particular database. This problem was recently tackled by [318] in the context of image classification with CNNs only. To the best of our knowledge, this work is the first to address this issue in the context of semantic features dealing with several tens of thousands of concepts.

4.2 Large scale semantic features

In this section, we detail our framework for building sparse semantic image descriptors from a large number of visual concepts, introduced in Chapter 3. We also present how we can exploit the obtained descriptors to represent a large image collection through an inverted index. This indexing choice leads to a significant acceleration of the retrieval process compared to direct search, as we show in Section 4.7. Our approach can be summed-up as: use an appropriate large-scale visual resource, represent concepts with linear models, exploit a good low-level feature, and sparsify to retain the most salient

dimensions. This approach is graphically illustrated in Figure 4.1. The focus is on the exploitation of automatically built resources, instantiated with Flickr groups, but the pipeline is generic and is also applied to ImageNet, a manually built dataset. The semantic descriptors obtained with these two resources are noted $Semfeat^{FG}$ and $Semfeat^{IN}$ respectively.

Individual visual models are used to map basic image features in a semantic space of size *sup* defined by the number of Flickr groups or ImageNet concepts used to build semantic features. We denote by $\mathbf{W}$ the matrix concatenating all individual visual models learned by (3.1):

$$\mathbf{W} = \{\mathbf{W}^1, ..., \mathbf{W}^{sup}\} \in \mathbb{R}^{(S_f+1)\times sup}. \tag{4.1}$$

Using the $\mathbf{W}$, an initial image feature $\mathbf{f} \in \mathbb{R}^{(S_f+1)}$ is mapped to its semantic representation $\mathbf{s} \in \mathbb{R}^{S_s}$ through

$$\mathbf{s} = \mathbf{W}^T\mathbf{f}. \tag{4.2}$$

$\mathbf{s}$ (a short notation for *Semfeat*) contains semantic information as it aggregates the classification scores of $\mathbf{f}$ given all available Flickr groups or ImageNet concepts. (4.2) is therefore comparable to a soft assignment encoding since individual classifiers contribute to the semantic representation.

In 4.3, we give an expanded view of 4.2. Each column of the leftmost matrix represent a classifier's learned weights and the bias term.

$$\begin{bmatrix} w_{1,1} & \cdots & w_{(S_f+1),1} \\ w_{1,2} & \cdots & w_{(S_f+1),2} \\ \vdots & & \vdots \\ w_{1,sup} & \cdots & w_{(S_f+1),sup} \end{bmatrix} \times \begin{bmatrix} f_1 \\ \vdots \\ f_{S_f} \end{bmatrix} = \begin{bmatrix} s_1 \\ \vdots \\ s_{sup} \end{bmatrix} \tag{4.3}$$

Semfeat is dense since all classifier outputs are taken into account. One drawback of such representation is that the effects of a relevant classifier can be smoothed by the accumulation of weights of an array of poorer classifiers. For instance, a concept whose output is 0.95 has lower importance than a combination of 5 concepts with 0.2 outputs. In order to circumvent this issue, we propose a sparsification method that is detailed below.

4.2.1 Semantic features sparsification

The authors of [343] showed that soft assignment encoding is not optimal as it discards the manifold geometric structure of the mapped space. Beyer et al. [344] demonstrated that distances between points often become less meaningful in high-dimensional space. Consequently, the expressive power of high-dimensional features is limited. Moreover, empirical analysis of high-dimensional features [345, 346] shows that they often lie on a manifold which has a much smaller intrinsic dimensionality. Such a manifold structure implies that the neighborhood of a feature point is homeomorphic to the Euclidean space into a local region only, thus computing distance (or proximity) between features is meaningful within a local region only. Outside of this region, two local points considered similar using a distance measure might actually be far from each other.

Given this manifold assumption, a classification score indicating the proximity of a feature $\mathbf{f}$ to a concept c is reliable only when $\mathbf{W}^c\mathbf{f}$ is large. The direct use of all concept classifiers degrades the semantic representation as concepts distant from $\mathbf{f}$ do not bring useful information. To obtain a more reliable representation, we leverage the manifold geometry by adding a locality constraint [346] to $\mathbf{s}$,

The locality constraint chooses the dimensions $\mathbf{s}$ based on the likelihood of the corresponding concepts with $\mathbf{f}$. Given this observation, an efficient sparse approximation of s can be derived. We define the matrix $\mathbf{W'}$ as follows:

$$\mathbf{W'} = \{\mathbf{W'}^1, ..., \mathbf{W'}^{S_s}\} \in \mathbb{R}^{(S_f+1)\times S_s} \tag{4.4}$$

which considers only the K individual visual models ($K << S_s$) that yield the highest scores on $\mathbf{f}$:

$$\mathbf{W'^c} = \begin{cases} W^c, \text{ if the} c^{th} \text{concept has one of the K highest prediction scores} \\ 0, \text{ otherwise} \end{cases} \tag{4.5}$$

We thus obtain an semantic representation which approximates locality through

$$\mathbf{s} = \mathbf{W'}^T\mathbf{f}. \tag{4.6}$$

To obtain a sparsified version of *Semfeat*, the only parameter is K which directly controls the locality of the solution.

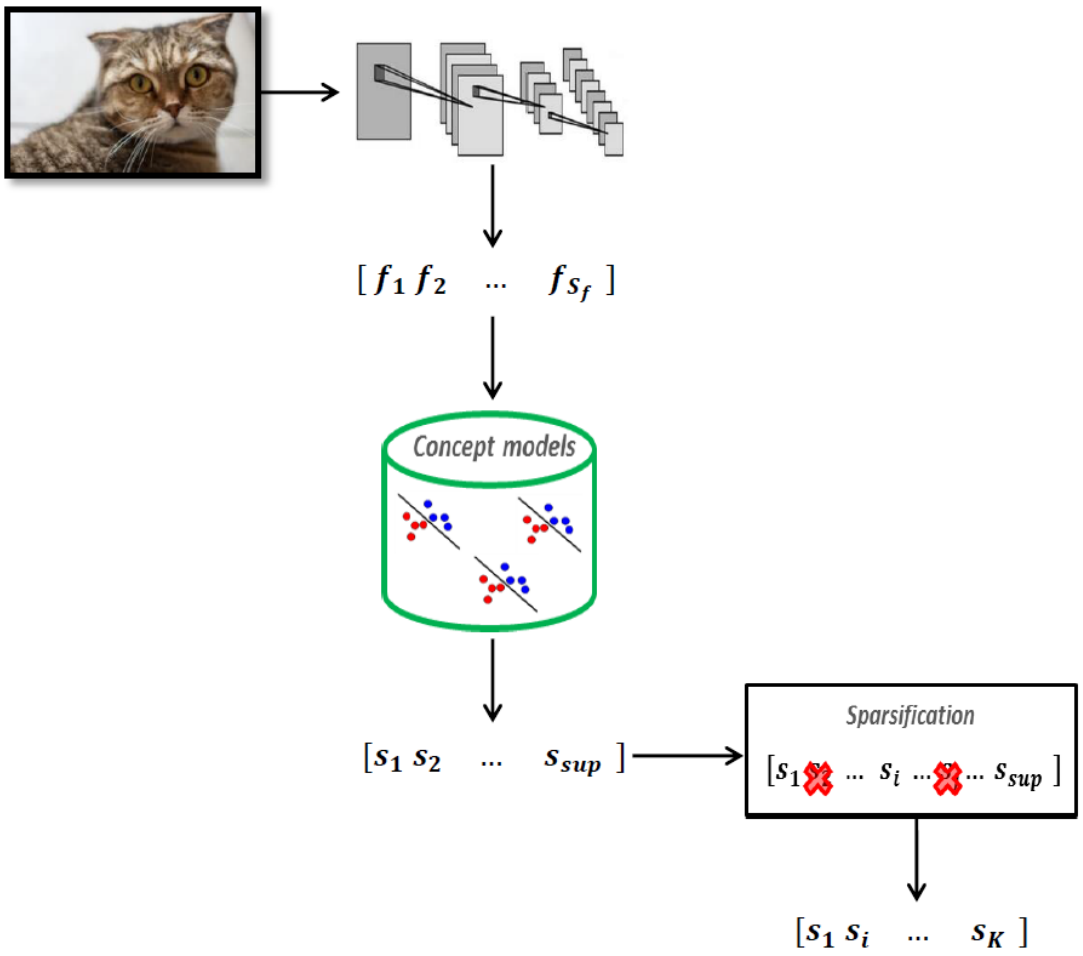
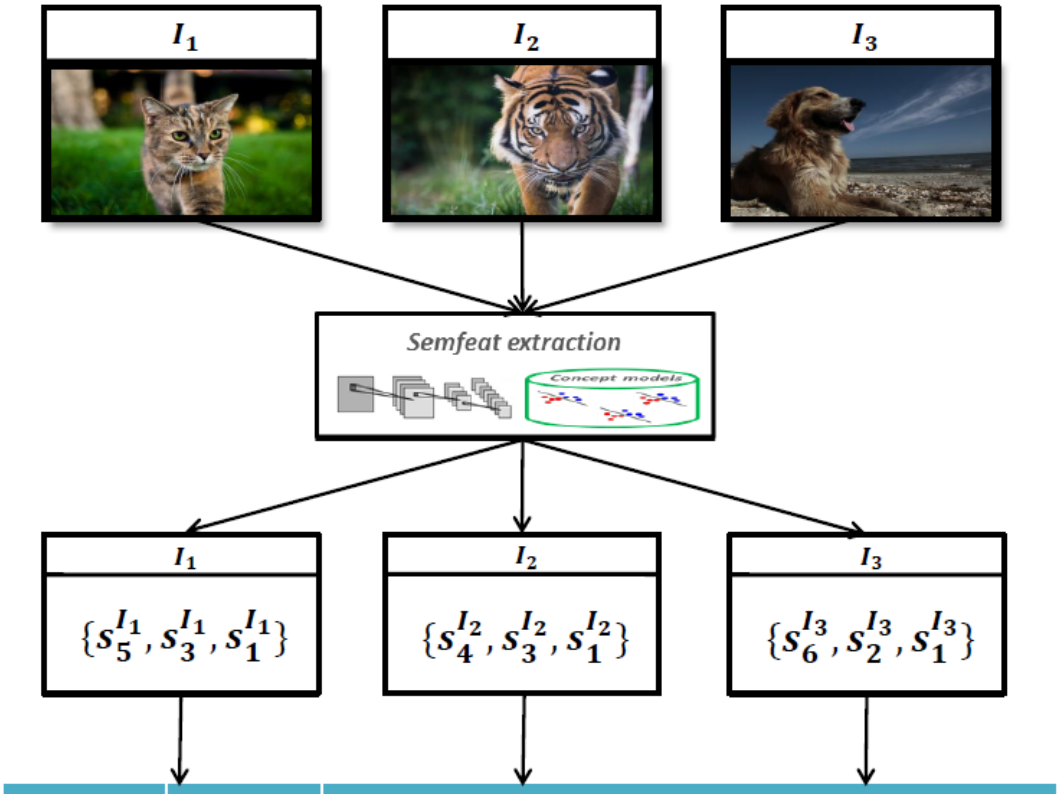


FIGURE 4.1: Illustration of the *Semfeat* extraction process.

In Figure 4.1, we illustrate the extraction and sparsification pipeline for obtaining the proposed semantic descriptor. For each image, we first extract the low-level descriptor. With the sole exception of Fisher [24] that is used in a preliminary experiments, all the initial image representation are CNN features (Overfeat, Caffe or VGG). These are detailed in chapter 3. In the following step, we use the matrix of learned concept weights (W) to obtain the dense *Semfeat* representation, as presented in Equation 4.3. W is obtained offline and the choice of training data and training configuration in presented in Chapter 3. This pipeline is independent of the initial feature used and the image collection used for training the visual concepts. We will test multiple configurations of semantic features for CBIR in Section 4.6. Finally, to obtain a compact representation of Semfeat, we sparsify the dense descriptor, as specified in Equation 4.6. We apply the same pipeline for all images in the collection and query images. Unless otherwise specified, we set the same value for the K parameter used in sparsification both for collection and query images.

4.2.2 Inverted index representation from semantic features



Concept Id	Concept name	Images (scores)
1	*mammal*	$(\,, s_1^{I_1})$ $(\,, s_1^{I_2})$ $(\,, s_1^{I_3})$
2	*canine*	$(\,, s_2^{I_3})$
3	*feline*	$(\,, s_3^{I_1})$ $(\,, s_3^{I_2})$
4	*tiger*	$(\,, s_4^{I_2})$
5	*cat*	$(\,, s_5^{I_1})$
6	*dog*	$(\,, s_6^{I_3})$

FIGURE 4.2: Simplified example of building an inverted index for an image collection from semantic features. In this particular case, we assume that after the sparsification stage, only 3 concepts are kept for each image. The score s_w^i represents the prediction score for image i of the visual model trained for concept w.

Inverted index structures have been extensively used for Web and text search for many years but also in image retrieval if the features are sparse [51]. This representation entails mapping each query word to a matching list of documents. The index servers then determine a set of relevant documents by intersecting the hit lists of the individual query

words, and they compute a relevance score for each document [138]. An inverted index data structure has been previously proposed for image retrieval. In one of the first works, Cao et al. [139] use binary counts of spatial bags of visual features based on SIFT descriptors. A more complex approach is detailed in [140]. There, a coding/decoding scheme used for the compression of tree-structured vector quantizer constructed by hierarchical k-means clustering of SURF descriptors.

Building an inverted index from semantic features is closer to the conventional use of inverted indexes for text search than previous attempts in CBIR. In our case, each component of the image descriptor conveys semantic meaning. We illustrate the process of building a semantic index for an image collection in Figure 4.2. For each image, we extract Semfeat and sparsify the feature, as detailed in the previous Section and in Figure 4.1. This entails that, for an image, only the top K most salient concepts are kept, along with their predicted scores. In our example, we assume $K = 3$. For each *activated* concept among the images from the collection, we build a dictionary with the concept id and the image ids, along with the prediction scores for all the images in which the concept is represented after sparsification. In contrast to the bag of visual words representation, each dimension of the inverted index conveys semantic meaning and can be used to give understandable feedback about image similarity to the user. Another advantage of using sparse semantic features is that compact representation (up to 200 non-null concepts) correlates with high retrieval performance, as we show in section 4.6. This entails that we do not need to use more complex and time consuming approaches (*e.g.* compression [140], hashing [347] or directly optimizing the classifier for binary features [4]) to obtain compact descriptors.

For retrieval, we rank the images found in the inverted index using the cosine similarity measure, written to exploit the sparse character of the features. If $I_q = \left\{s_1^q, s_2^q, \cdots, s_K^q\right\}$ is the query image, we define the similarity between I^q and an image I^c from the inverted index as follows:

$$\mathbf{sim_{cos^{iv}}}\left(\mathbf{I_q}, \mathbf{I_c}\right) = \frac{\sum_{s_i^{I_q} \in I^q} s_i^{I_q} \cdot s_i^{I_c}}{\|I_q\| \, \|I_c\|} \tag{4.7}$$

In Equation 4.7, I_c is a collection image that appears in the list of images in the inverted index for at least one of the concepts present in the query image and $\|I_c\|$ is the Euclidean norm of the *Semfeat* descriptor of image I_c. In practice, for each concept *activated* in I_q, we first retrieve from the inverted index only the list of collection its associated images and adapt their similarity score $sim_{cos_{iv}}\left(I_q, I_c\right)$. This approach increases the retrieval speed by not having to compute the similarity between the query image and all the images in the collection. However, one potential drawback, especially for small scale collections, is that the retrieved number of images might be insufficient. The size of the

result list depends on the level of sparsification K. We investigate the impact of the number of retained concepts in Section 4.4.

4.3 Experimental setup

In this Section, we first describe the datasets used for evaluating the retrieval performance of our proposed semantic features. We also present six existing feature that we use as baselines in one or several experiments.

4.3.1 Evaluation datasets

The main objective here is to assess the usefulness of *Semfeat* in a CBIR task performed over diversified datasets. We are also interested in evaluating different configurations of *Semfeat*, both in terms of initial feature representation, data source used for building visual concept models, conceptual coverage and sparsification level. Next, we present three datasets used in our tests and the experimental setup fixed for each one.

Wikipedia Retrieval 2010 Wikipedia Retrieval 2010 was created as part of the ImageCLEF evaluation campaign[1] and is publicly available. It includes 237,434 Wikimedia images which were extracted from a large and diversified subset of Wikipedia articles and is publicly available. This collection it thus fitted for ad-hoc image retrieval experiments, in which any query can be submitted to the process. To ensure comparability with other methods already tested on this dataset, we report mean average precision (MAP) performances. The 2010 CBIR query set contains 118 query images for 70 diversified queries. CBIR over the Wikipedia collection is challenging because the image content is highly diversified. In addition, some topics are represented by only few relevant images. The Wikipedia Retrieval ground truth has been built using a pooling approach and is therefore incomplete [348].

To improve comparability of existing and new runs, we extend the original ground truth (noted $origGT$) by pooling the new runs proposed here. This extension (noted $extGT$) is realized using similar topic narratives and a majority voting with three relevance judgments per image. We had limited resources and assessed only the new images appearing in the top 20 results of a selection of runs are annotated, compared to a pooling depth of 100 used for establishing the initial ground truth. Experiments are performed using the 118 example images associated with 70 topics of the Wikipedia Retrieval dataset. If

[1] http://www.imageclef.org/

there are two examples per topic and a collection image is similar to both of them, the highest similarity score is retained.

MIRFLICKR The MIRFLICKR-25000[2] dataset [349] consists of images retrieved from Flickr along with their user assigned tags and it was used in the ImageCLEF 2009 and ImageCLEF 2010 Photo Annotation tasks. 25,000 have been annotated for 24 topics including object categories (*e.g. bird, tree, people*) and scene categories (*e.g. sky, indoor, night*). A stricter labeling was done for another 14 classes where an image was annotated with a category only if that category was salient. This leads to a total of 38 classes that are labeled for all the images of the collection. Similar to Beecks et al. [350], we randomly take 1000 images as queries. We use the rest of the 24,000 images as our test collection. We ensure that each of the 38 topics is represented at least 5 times among the queries. As an image may be labeled with multiple concepts, each query image is used in retrieval tests for all of the associated concepts.

NUS-WIDE NUS-WIDE[3] [8] is a large scale dataset collected from Flickr. It contains 269,648 images, provided as multiple visual features and source URLs, with 5,018 tags of which 81 have been manually checked and can be considered ground-truth tags. Unfortunately, some images are not available anymore, therefore we had to use a subset of 205,347 images that were still present on Flickr at the time we downloaded the images. We use an experimental set-up inspired by Wang et al. [351] and Tang et al. [352] but increase the query set size from 1000 to 5000 to improve statistical significance. Similar to the MIRFLICKR evaluation setup, we ensure that each of the 81 concepts has at least 5 query images.

4.3.2 Image representations

The features described in this Section are used in one or several experiments detailed in this Chapter.

Fisher Vectors, Overfeat, Caffe and VGG are strong baselines, representing both pre-CNN and CNN features:

- *Fisher* - existing baseline which exploits a version of Fisher Vectors adapted for CBIR [24].

[2] http://press.liacs.nl/mirflickr/
[3] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

- *Overfeat* - baseline which exploits the default outputs of the CNN-based feature extractor presented in [14], using the small network to speed-up feature computation. For more details, see Section 3.2.2.

- *Caffe* - Image representation resulted from the standard model provided by Caffe (*i.e.* the original model of Krizhevsky et al. [63]). Similar to Overfeat, we extract the weights of the *fc7* layer, followed by a L2 normalization.

- *VGG* We will refer by *VGG* the Caffe models of the networks described in [65]. For more details, see Section 3.2.2.

We also want to compare Semfeat to existing semantic image descriptors that are publicly available:

- *PiCoDes* - PiCoDes (Picture Codes) [4] use basis classifiers as features with linear models. citetbergamo2011picodes learn abstract categories aimed at optimizing linear classification when they are used as features. This learning objective decouples the number of training classes from the target dimensionality of the binary descriptor and thus it allows the optimization of the descriptor for any arbitrary length. The learned features describe the image in terms of binary visually-interpretable properties corresponding, e.g., to particular shape, texture or color patterns. We extract the largest version of PiCoDes, which represents a vector of 2048 binary features.

- *MC* - The Meta-class [23] descriptor is obtained through a hierarchically partition over the set of training object classes such that each meta-class subset can be easily recognized from the others. This criterion forces the classifiers trained on the meta-classes to be repeatable. The meta-classes are superclasses of the original training categories and capture common visual properties shared by similar classes. The MC descriptor has a size of 15,232 dimensions.

PiCoDes and Met-classes are computed with the VLG extraction tool[4]. Although it offers the possibility to also extract other features, we chose to use PiCoDes and Meta-class as our baseline semantic features due to their superior performance on standard datasets [5].

Next, we present the different configurations of the proposed semantic descriptors (*Semfeat*) that are evaluated throughout this Chapter:

[4]http://vlg.cs.dartmouth.edu/projects/vlg_extractor/vlg_extractor/Home.html

- $SemfFeat^{IN}_{Overfeat}$ - semantic feature based on the 17,462 ImageNet concepts which were introduced in Chapter 3. The visual models are built using Overfeat features. Results are reported for sparsification $K = 10$, corresponding to the best $Semfeat^{IN}$ MAP in figure 4.3.

- $Semfeat^{FG}_{Overfeat}$ - semantic feature based on the 30,000 groups with the highest cross-validation scores from the initial 38,500 group dataset. The visual models are built using Overfeat features. The sparsification is set to $K = 30$.

- $SemfFeat^{IN}_{Caffe}$ - semantic feature based on ImageNet concepts using visual models are built upon Caffe features. Unless otherwise specified, the sparsification level is set to $K = 20$.

- $SemfFeat^{FG}_{Caffe}$ - semantic feature based on Flickr groupsusing visual models are built upon Caffe features. Unless otherwise specified, the sparsification level is set to $K = 30$.

- $SemfFeat^{IN}_{VGG}$ - semantic feature based on ImageNet concepts using visual models are built upon VGG features. Unless otherwise specified, the sparsification level is set to $K = 20$.

- $SemfFeat^{FG}_{VGG}$ - semantic feature based on Flickr groupsusing visual models are built upon VGG features. Unless otherwise specified, the sparsification level is set to $K = 30$.

As shown in Chapter 3, we benefit from a larger number of negative training samples when training individual models. In preliminary experiments on the Wikipedia Retrieval 2010 dataset, we test two approaches for choosing negative training instances. The first one uses fixed set of 100,000 negatives. In the second one, we dynamically choose the number of negatives from a large class, proportionally to the number of negatives (*i.e.* n times the number of negatives as there are positives, with $n \in \{1, 10, 100\}$). For the Overfeat based *Semfeat* configurations tested in this Chapter, we keep the fixed number of 100,000 negatives. For the Caffe and VGG based ones, we found that a ratio of $n = 100$ is best suited for retrieval and it will be used for all $SemfFeat_{Caffe}$ and $SemfFeat_{VGG}$ detailed in this chapter. For instance, $SemfFeat^{IN}_{Caffe}$ the MAP score on Wikipedia Retrieval 2010 increases from 0.1258 to 0.1446 and to 0.1546 on the original ground-truth, when increasing n from 1 to 10 and 100, respectively.

4.4 Sparsification evaluation

One of our central objectives is to create features which are in the same time efficient and compact. To this end, sparsity is a desirable property of document representations since it allows one to use inverted indexes in order to search massive datasets in real time. Sparsification is applied to semantic features built on top of 30,000 Flickr Groups and of ImageNet, as detailed in Subsection 4.2.1 Those datasets have respectively 30,000 and 17,462 concepts available in each case.

4.4.1 Collection specific sparsification

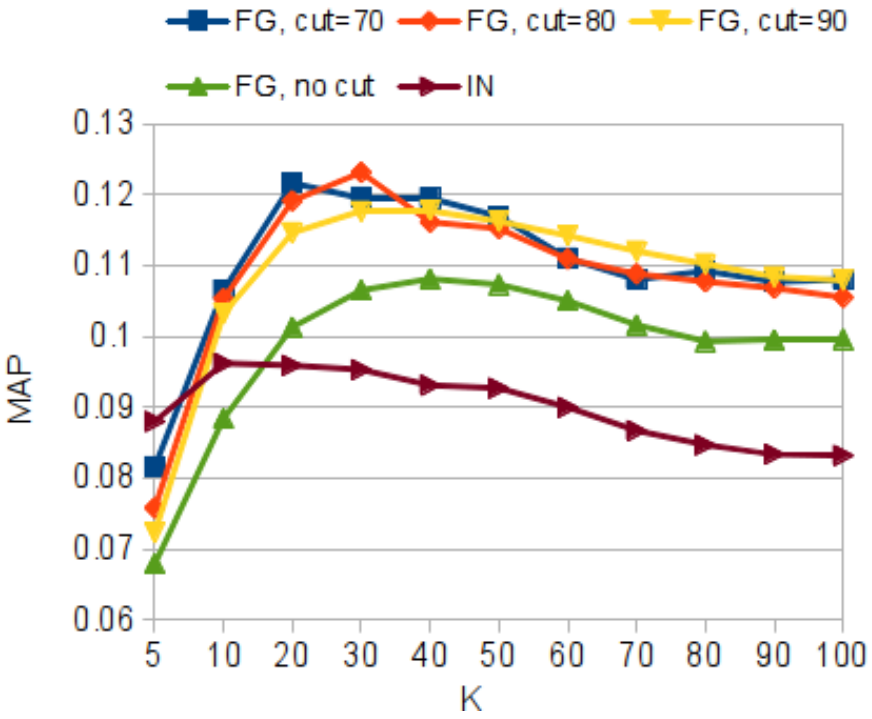


FIGURE 4.3: Sparsification analysis in function of K, the number of most salient concepts retained in the semantic features built on top of Flickr groups and of ImageNet on the Wikipedia Retrieval dataset. The models are trained with Overfeat features.

In a first experiment, we compare the retrieval performance (measured by the MAP score) of $SemfFeat^{IN}_{Overfeat}$ and $SemfFeat^{FG}_{Overfeat}$ features on the Wikipedia Retrieval 2010 collection. We are also interested to compare Flickr group based visual models trained using different percentages of ranked images for positive instances. We use the *skNN* method introduced in Chapter3 for ranking. We note by $Semfeat^{FG}_{Overfeat_{cut}}$ the *Semfeat* obtained from model train using only the first *cut%* ranked positive images. In figure 4.3, we vary the sparsification factor $K = \{5, 10, ..., 100\}$ in order to thoroughly evaluate the effect of sparsification. Very interestingly, all group based features compare favorably

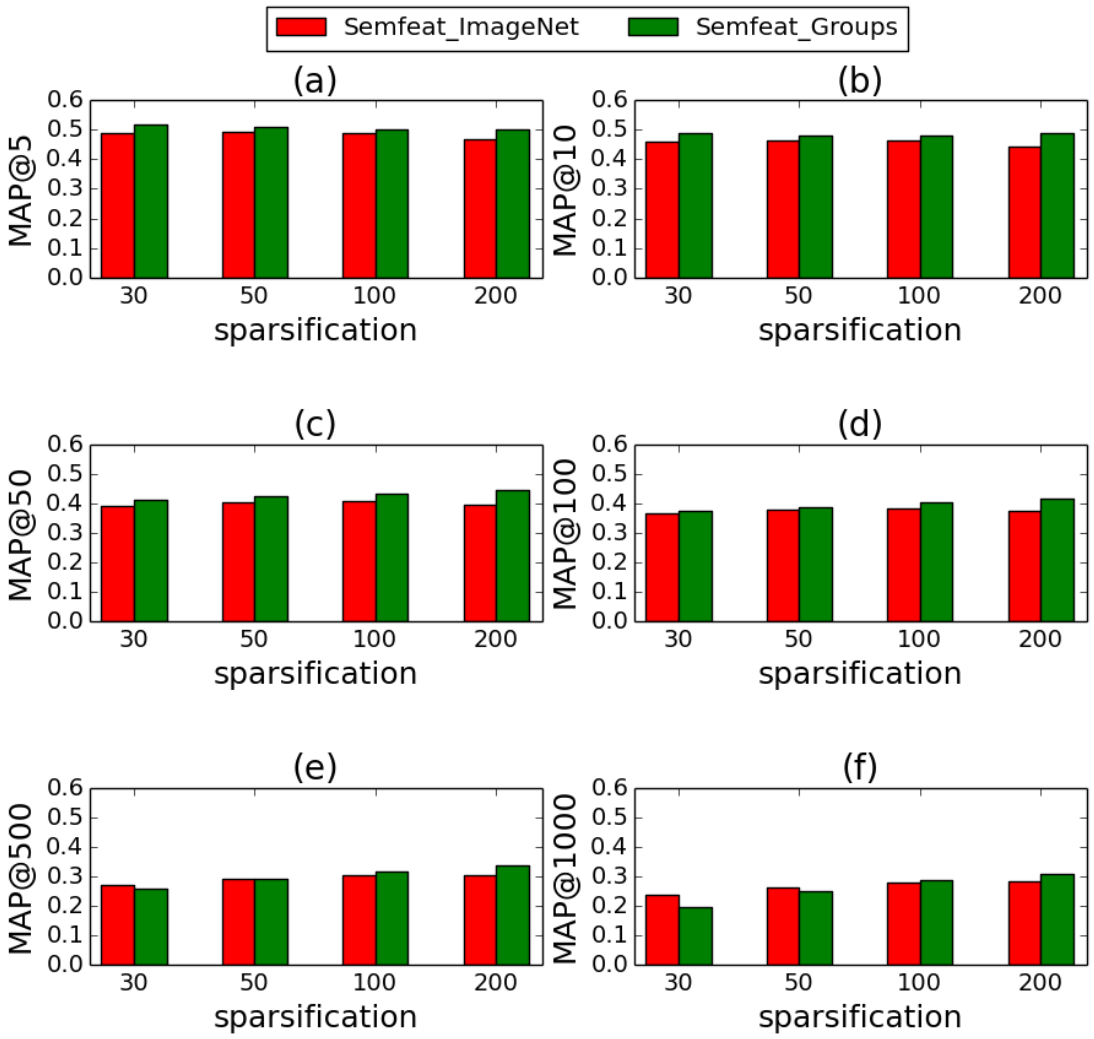


FIGURE 4.4: Retrieval results for $Semfeat^{IN}_{VGG}$ and $Semfeat^{FG}_{VGG}$ features with 4 sparsification levels ($K = \{30, 50, 100, 200\}$) on the MIRFLICKR dataset. In subplot *(a)* we evaluate with MAP@5, in subplot *(b)* with MAP@10, in subplot *(c)* with MAP@50, in subplot *(d)* with MAP@100, in subplot *(e)* with MAP@500, and in subplot *(f)* we evaluate with with MAP@1000.

with *Semfeat* for $K > 10$. This finding confirms that Flick groups can be successfully used in CBIR as a substitute for manually built resources.

The results presented in figure 4.3 also show that the most interesting MAPs are obtained when 10 to 50 most salient concepts detected in them are used in the representation, with small peaks around $K = 30$. This finding has an important practical implication since *Semfeat* features can be efficiently represented using inverted indexes. As a result, it is possible to search very large image datasets much faster than with a brute force search which is needed for most existing low-level or intermediate features, such as Fisher Kernels or Overfeat. Confirming the results presented in table 3.2, the performances obtained with cut-offs $cut = \{70, 80, 90\}$ are close to each other. The constant gap

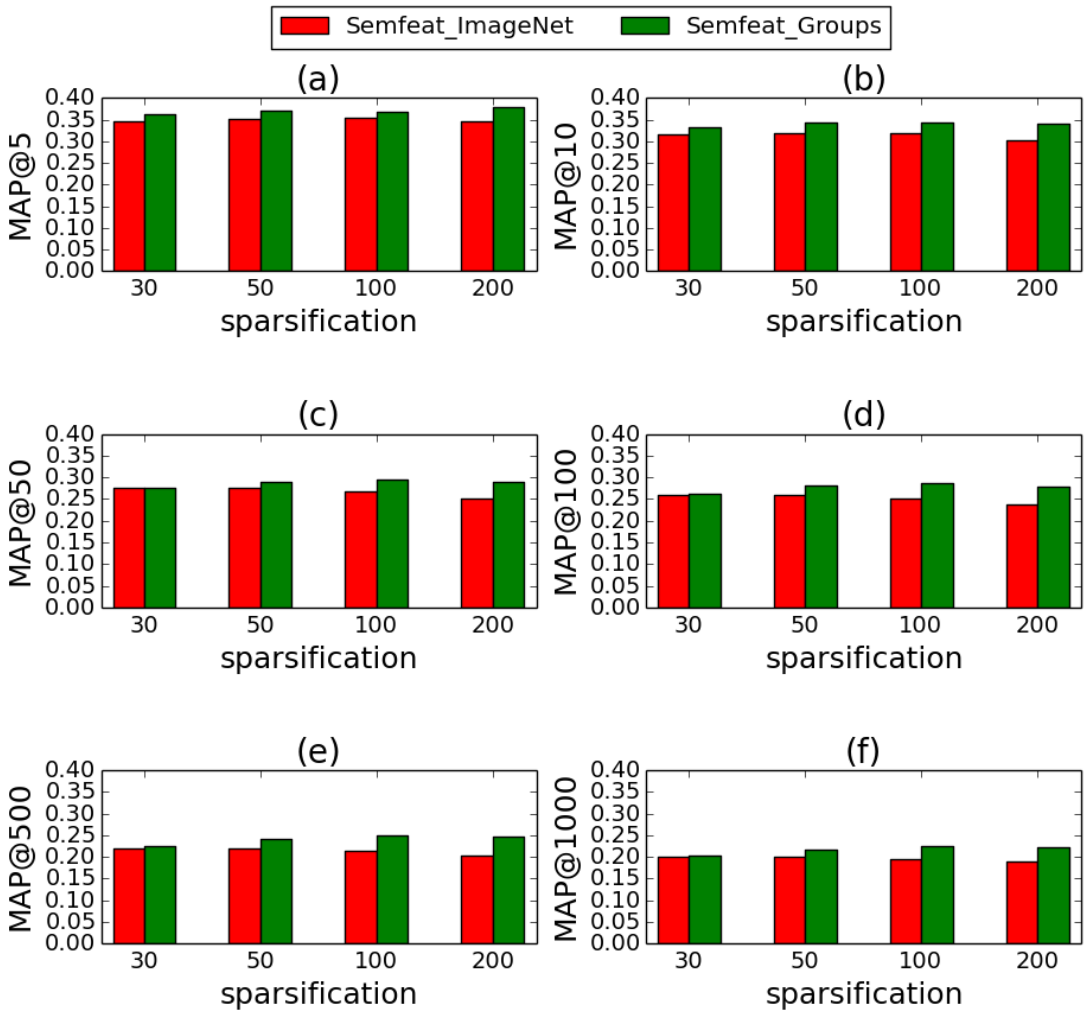


FIGURE 4.5: Retrieval results for $Semfeat^{IN}_{VGG}$ and $Semfeat^{FG}_{VGG}$ features with 4 sparsification levels ($K = \{30, 50, 100, 200\}$) on the NUS-WIDE dataset. In subplot *(a)* we evaluate with MAP@5, in subplot *(b)* with MAP@10, in subplot *(c)* with MAP@50, in subplot *(d)* with MAP@100, in subplot *(e)* with MAP@500, and in subplot *(f)* we evaluate with with MAP@1000.

between reranked versions of $Semfeat^{FG}_{Overfeat_{cut}}$ and $Semfeat^{FG}_{100}$ indicates that removing poorly reranked images has a clear beneficial effect regardless of the sparsification factor. Results obtained with the three *cut* values are rather similar, indicating that the positive effect of noise reduction and the negative effect due to training set shrinking compensate each other. Beyond $K = 100$, performances drop continuously, an effect which is mainly explained by the fact that image similarities are computed based on visual concepts which are loosely associated to image content.

For all of the remaining experiments, we fix *cut* at 80. We will simply refer as $Semfeat^{FG}_{Feature}$ the *Semfeat* descriptor obtained from models trained using 80% of positive instances.

In the following experiments, we compare $Semfeat^{IN}_{VGG}$ and $Semfeat^{FG}_{VGG}$ features with 4

sparsification levels ($K = \{30, 50, 100, 200\}$) both on the MIRFLICKR dataset (in Figure 4.6) and NUS-WIDE (in Figure 4.5). We test the precision of the retrieved results at different levels and we evaluate with MAP@k, with $k \in \{5, 10, 50, 100, 500, 1000\}$. We first observe in Figure 4.5 that the *Semfeat* features built upon Flickr groups outperform those based on ImageNet concepts for all sparsification levels (K) and for all evaluation metrics. The difference is more noticeable with the increase of K. This result is consistent with the finding for Wikipedia Retrieval 2010, presented in Figure 4.3. *Semfeat* descriptors built from Flickr groups perform better when more concepts are kept in the descriptor than ImageNet based *Semfeat* representations. The same conclusion can be drawn when testing on the MIRFLICKR dataset (Figure 4.6). There, the only exception can be found in subplot *(f)*, where, at $K = 50$, $Semfeat^{IN}_{VGG}$ surpasses $Semfeat^{FG}_{VGG}$. For both datasets, but more evident in the case of MIRFLICKR, guarding more concepts in *Semfeat* increases the MAP@500 and MAP@1000 scores. This observation differs from what can be deducted from Figure 4.3, where setting the sparsification level K over 30 reduces the MAP score for all *Semfeat* configurations. One possible explanation of this dataset bias is that, due to the inverted index structure that is used for retrieval, the number of retrieved images is influenced by the sparsification level (see Subsection 4.2.2). When there are a large number of relevant images for most queries, such is the case of MIRFLICKR and NUS-WIDE, having a low K leads to low recall and it reflects on the MAP@k score, when $k > 100$. On the contrary, for the Wikipedia Retrieval dataset, both for the original and the extended ground-truths, we have a lower number of relevant images per query.

4.4.2 Query image sparsification

In order to ensure coherence in the extraction of semantic descriptors, in most experiments we use the same sparsification level for collection and query images. However, in this Section, we investigate the effect of varying the number of concepts kept in the query image, while having an inverted index representation for the image collection obtained from a fixed number of concepts. For the results depicted in Figure 4.6, we set the sparsification level for the collection images at $K = 200$. We evaluate on MIRFLICKR and compare the retrieval performance of $Semfeat^{IN}_{VGG}$ and $Semfeat^{FG}_{VGG}$, with 20 sparsification levels for the query images ($K_{query} = \{10, 20, ..., 200\}$). The only evaluation configuration where both $Semfeat^{IN}_{VGG}$ and $Semfeat^{FG}_{VGG}$ display a similar behavior when increasing the number of concepts in the query is when looking only at the top 5 results (subplot *(a)*). In this case, the highest MAP score is obtained for $K_{query} = 20$ for $Semfeat^{FG}_{VGG}$ and $K_{query} = 30$ for $Semfeat^{IN}_{VGG}$. These values are similar to the sparsification setting used for the Wikipedia Retrieval dataset (Figure 4.3).

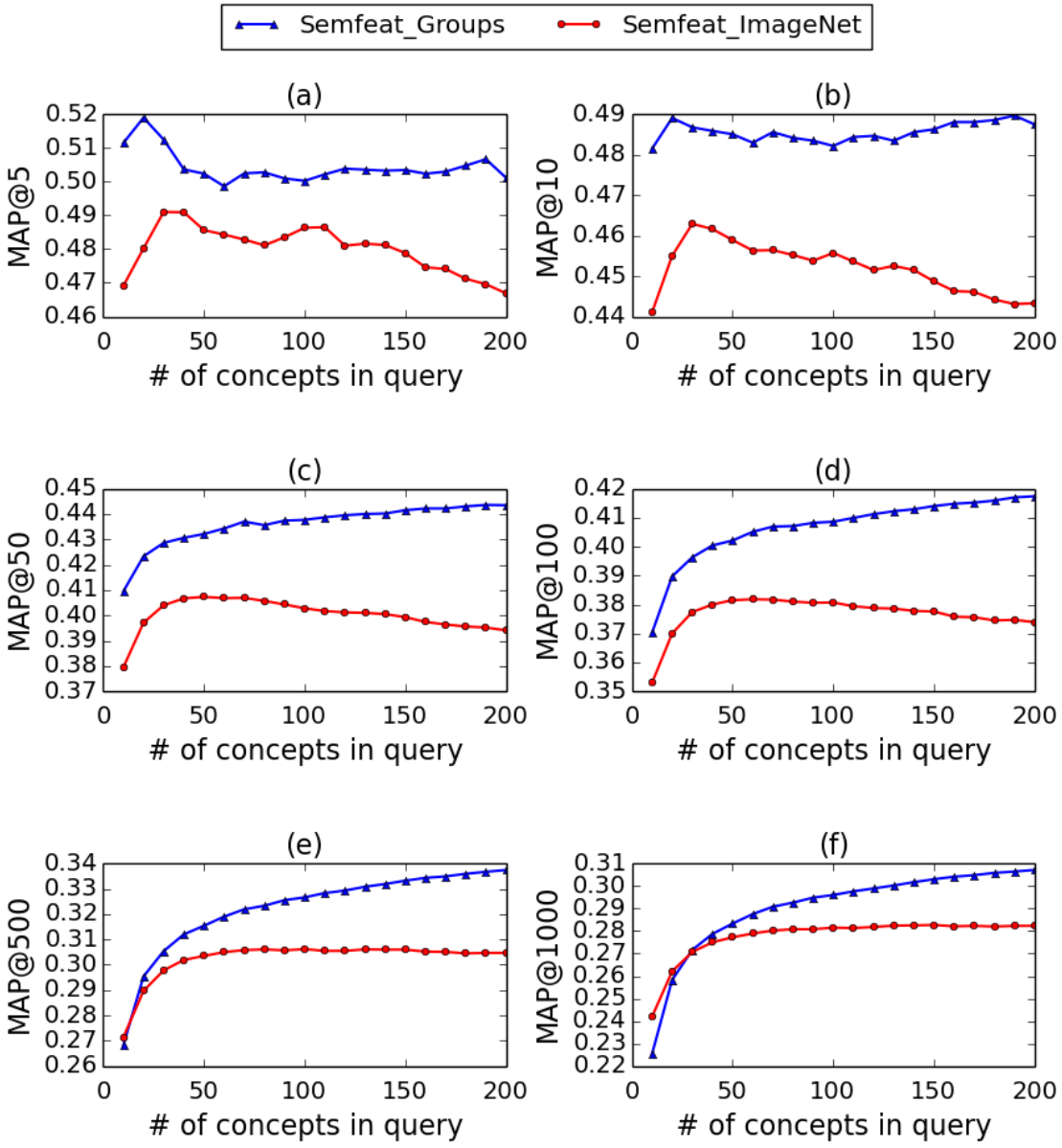


FIGURE 4.6: Retrieval results for $Semfeat^{IN}_{VGG}$ and $Semfeat^{FG}_{VGG}$ features with 20 sparsification levels ($K = \{10, 20, ..., 200\}$) for query images on the MIRFLICKR dataset. For the collection images, the sparsification level is set at $K = 200$. In subplot *(a)* we evaluate with MAP@5, in subplot *(b)* with MAP@10, in subplot *(c)* with MAP@50, in subplot *(d)* with MAP@100, in subplot *(e)* with MAP@500, and in subplot *(f)* we evaluate with with MAP@1000.

For $Semfeat^{IN}_{VGG}$, increasing K_{query} from 40 has a strong negative impact on MAP@5 and MAP@10 and leads to a slight drop of the MAP@50 and MAP@100 scores. For MAP@500 and MAP@1000 (subplots *(e)* and *(f)*), we can observe a plateau after $K_{query} = 50$. On the contrary, in the case of $Semfeat^{FG}_{VGG}$, having more concepts kept in the query images has a positive effect when looking at the top 50, 100, 500 and 1000 retrieved images. This may be explained both by the lower number of support concepts

that make up $Semfeat^{IN}$, compared to $Semfeat^{FG}$. This result confirms that having a large number of concepts for building *Semfeat*, coupled to keeping more concepts after sparsification is beneficial for retrieving a higher set of images. We also reinforce the observation that using Web images for visual concept training outperforms a manually curated resource in our CBIR framework.

4.5 Conceptual coverage evaluation

The richness of the conceptual support of semantic features is tightly linked with their capacity to deal with heterogeneous datasets and we evaluate the influence of the size of this support. We investigate in this Section the impact of the number of concepts used for building *Semfeat* before sparsification. In a first experiment, we randomly select subsets of ImageNet concepts from the full collection. Next, we investigate another approach of selecting concepts, by taking into account the cross-validation accuracy ranking of the support visual concept models.

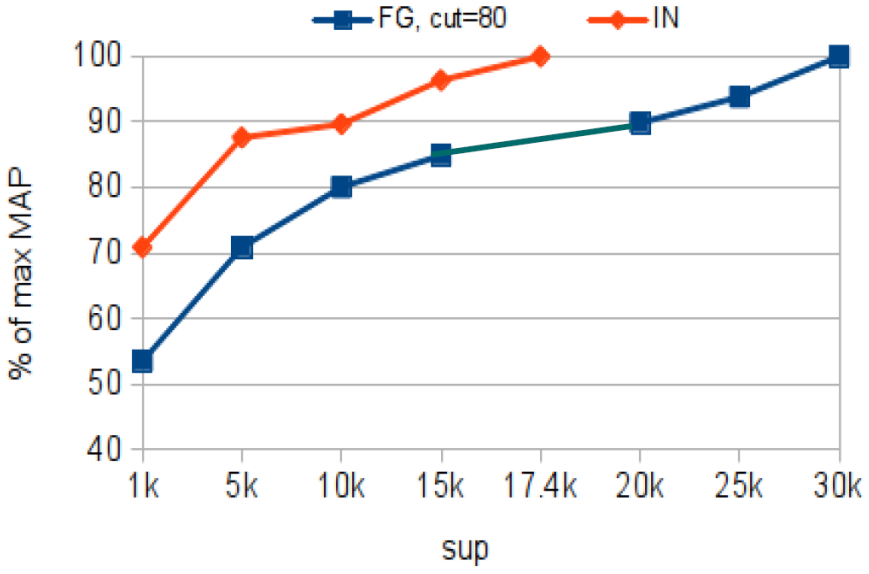


FIGURE 4.7: Conceptual coverage influence on the Wikipedia Retrieval dataset for $Semfeat^{IN}_{Overfeat}$ and $Semfeat^{FG}_{Overfeat}$. Results are reported by the percentage of the best MAP score (when using all concept classifiers) obtained by each configuration.

For the results presented in figure 4.7, we fix the *skNN* reranking percentage $cut = 80$ and sparsification at $K = 30$ and vary the conceptual support (*sup*) between 1000 and 30,000 for $Semfeat^{FG}_{Overfeat}$ and from 1000 to 17,462 for $Semfeat^{IN}_{Overfeat}$. To simulate subsets of the available resources, the groups included in each support are selected randomly. The results show that the number of support concepts has a non-negligible influence on semantic features performance. For instance, the MAP obtained with a support of 5,000

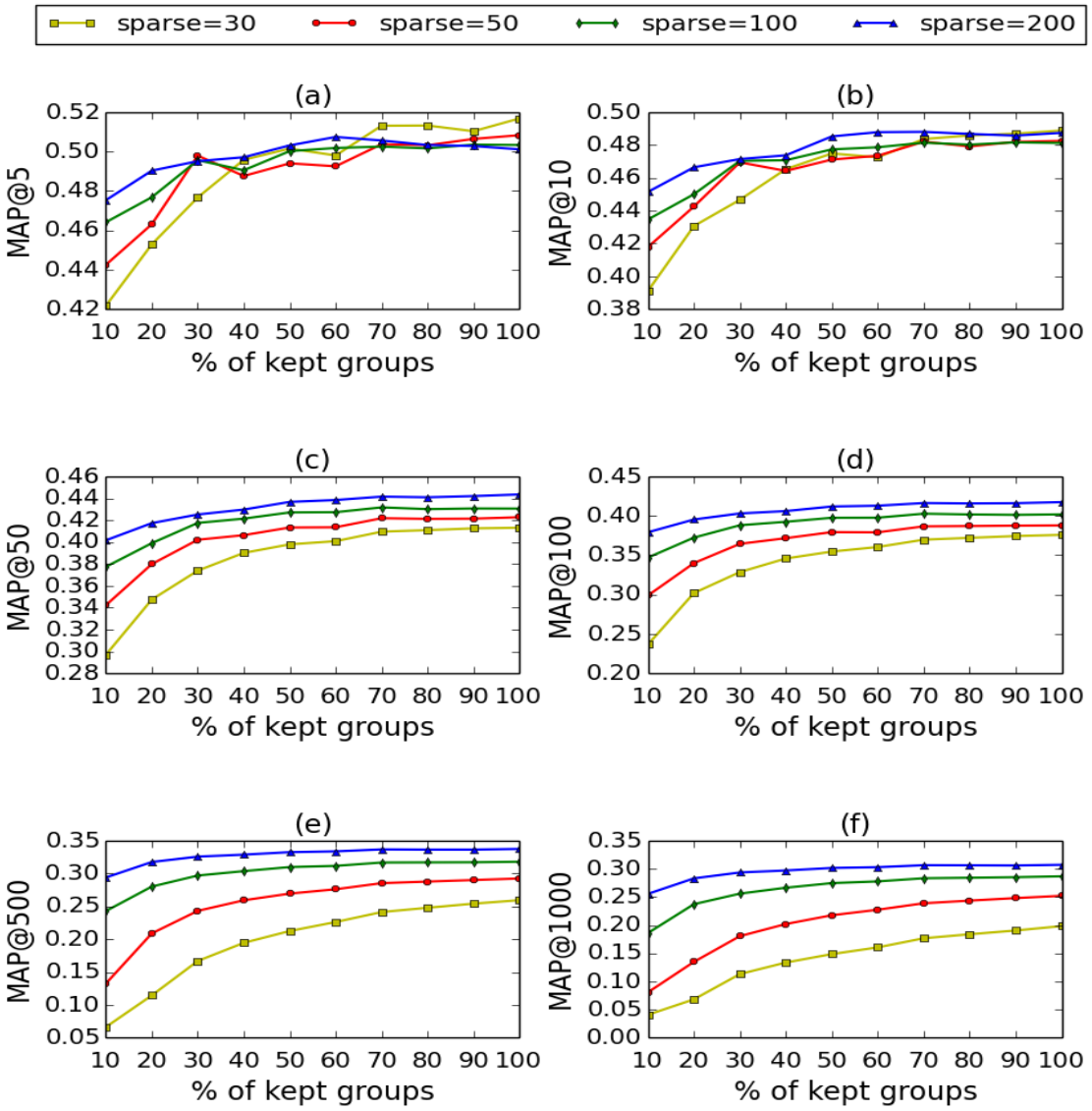


FIGURE 4.8: Retrieval results for $Semfeat^{FG}_{VGG}$ on the MIRFLICKR dataset when using only the top $n\%$ of Flickr groups from the list of groups ranked according to the cross-validation scores. We compare 4 sparsification levels ($K = \{30, 50, 100, 200\}$). In subplot *(a)* we evaluate with MAP@5, in subplot *(b)* with MAP@10, in subplot *(c)* with MAP@50, in subplot *(d)* with MAP@100, in subplot *(e)* with MAP@500, and in subplot *(f)* we evaluate with with MAP@1000.

concepts is 70% of the maximum MAP for $Semfeat^{FG}_{Overfeat}$ and 88% for $Semfeat^{IN}_{Overfeat}$. This effect of size is more important on $Semfeat^{FG}_{Overfeat}$ a behavior which is probably explained by the fact that automatically selected groups are more redundant than ImageNet concepts. The slope is lower for high values of *sup* but saturation is not reached in either case. This indicates that adding supplementary concepts or groups would probably have a positive effect on performances.

In Figure 4.8, we investigate the impact of the number of initial visual concepts used in

Semfeat. We test with $Semfeat^{FG}_{VGG}$ on the MIRFLICKR dataset. We first rank the Flickr groups according to their cross-validation scores. We then build *Semfeat* descriptors from 10 subsets of groups. We test with the top ranked $n\%$ groups, with n ranging from *10* to *100*. We already know from the experiments presented in Section 4.4 that using Flickr groups as base visual concepts and a high K for sparsification provides the highest MAP scores on MIRFLICKR. In the experiment depicted in this Figure, we chose to compare 4 sparsification levels ($K = \{30, 50, 100, 200\}$). Similar to the finding for the Wikipedia Retrieval dataset depicted in Figure 4.7, when using lower sparsification values ($K = 30$ and $K = 50$), the MAP@k scores monotonically increase when using more groups for most of k values. The most noticeable correlation between the retrieval performance and the MAP scores is for MAP@500 and MAP@1000 (subplots *(e)* and *(f)*). On the contrary, when applying a high sparsification value on *Semfeat*, the number of initial visual concepts is less important. For $K = 200$, there is a less than 0.01% points increase in MAP@50 and MAP@100 scores when using 100% of the groups instead of the first 60%. Moreover, we obtain the same MAP@500 and MAP@1000 when including 50% of the groups (15,000) in *Semfeat* as in the setting in which we include all 30,000 groups. This result entails that with proper visual concept ranking and a higher sparsification level, we near-optimal retrieval performance when halving the number of support visual concepts in *Semfeat*.

4.6 CBIR results

In this Section, we present an overview of the best configurations for *Semfeat* on the 3 evaluation datasets introduced in Subsection 4.3.1. As we mentioned, *Semfeat* is compared to strong pre-CNN and CNN descriptors, as well as state-of-the art semantic image descriptors. We also go into a deeper analysis of the results on all datasets to see on which topics the semantic features introduced in this Chapter stand out and what are the queries for which they underachieve.

4.6.1 Results overview

In table 4.1, we compare the MAP scores obtained with *Fisher* and *Overfeat* baselines to those with *Semfeat* versions proposed in this chapter. The Fisher vector is dense and contains over 100,000 dimensions [24]. The Overfeat vector is also dense and is obtained with the small network configuration [14]. For $Semfeat^{IN}$, we use a sparsification factor $K = 20$, which corresponds to the best MAP reported in Figure 4.3. For $Semfeat^{FG}$, results are reported for sparsification $K = 30$ and $cut = 80$, corresponding to the best MAP in Figure 4.3.

TABLE 4.1: Results for CBIR runs with the ImageCLEF Wikipedia Retrieval 2010 dataset. Both the original and the extended ground truth (*origGT* and *extGT*) are used. $Semfeat^{FG}$ results are reported for sparsification $K = 30$. *Fisher* performances do not change since this run was already pooled during the creation of *origGT*.

	MAP *origGT*	MAP *extGT*
Fisher	0.0553	0.0553
Overfeat	0.0986	0.1149
Caffe	0.1259	0.1373
VGG	0.1683	0.1849
$Semfeat^{IN}_{Overfeat}$	0.0962	0.1167
$Semfeat^{FG}_{Overfeat}$	0.1065	0.1267
$Semfeat^{IN}_{Caffe}$	0.1546	0.1658
$Semfeat^{FG}_{Caffe}$	0.1696	0.1837
$Semfeat^{IN}_{VGG}$	0.1955	0.2087
$Semfeat^{FG}_{VGG}$	**0.2127**	**0.2276**

The results presented in Table 4.1 show that $Semfeat^{IN}_{Overfeat}$ has roughly the same performances as *Overfeat* while $Semfeat^{FG}_{Overfeat}$ is better than this strong baseline. However, $Semfeat^{FG}_{VGG}$ gives a 26.3% and 23% relative improvement compared to *VGG* with *origGT* and *extGT*. Also , $Semfeat^{FG}_{Caffe}$ gives a 33.6% and 33.7% relative improvement compared to *Caffe* with *origGT* and *extGT*. When comparing $Semfeat^{FG}_{VGG}$ with $Semfeat^{IN}_{VGG}$, we see a 8.7% relative improvement on *origGT* and a 9% relative improvement on *extGT*. Similarly, when comparing $Semfeat^{FG}_{Caffe}$ with $Caffe^{IN}_{VGG}$, we see a 9.7% relative improvement on *origGT* and a 10.7% relative improvement on *extGT*. Compared to *Fisher*, the previous state-of-the-art method tested on this dataset, improvements are very consequent, 122.6% and 169% relative improvements for $Semfeat^{FG}_{Overfeat}$ for the two ground truths.

The best run is obtained with $Semfeat^{FG}_{VGG}$ ($MAP = 0.2127$). For comparison, the best text run submitted during the ImageCLEF campaign, which combined annotations in different languages and sophisticated language models, had $MAP = 0.2361$ [24]. From the results presented in this table, we can conclude that the CNN features used for training the visual models that are part of *Semfeat* directly influence the performance of the semantic descriptor. We also notice a 300% relative improvement with the best CNN feature (VGG) over the Fisher descriptors.

In Table 4.2, we present an overview of results for CBIR runs on the MIRFLICKR dataset and in Table 4.3, we present an overview of results for CBIR runs on the NUS-WIDE dataset. From the initial experiments on the Wikipedia Retrieval 2010 dataset (Table 4.1), we see that CNN features clearly surpass Fisher descriptors and, among CNN features, Caffe and VGG outperform Overfeat descriptors. Also $Semfeat_{Caffe}$ and $Semfeat_{VGG}$ are better than $Semfeat_{Overfeat}$ both for ImageNet and Flickr groups for a

TABLE 4.2: Results for CBIR runs with the MIRFLICKR dataset. $Semfeat^{IN}$ and $Semfeat^{FG}$ results are reported for sparsification $K = 200$ and query sparsification $K_{query} = 200$.

	Metric: MAP@k					
	k=5	*k=10*	*k=50*	*k=100*	*k=500*	*k=1000*
Caffe	0.437	0.397	0.327	0.300	0.230	0.207
VGG	**0.479**	**0.449**	0.379	0.347	0.267	0.240
MC	0.305	0.265	0.211	0.191	0.150	0.135
PiCoDes	0.283	0.249	0.196	0.178	0.134	0.118
$Semfeat^{IN}_{Caffe}$	0.424	0.405	0.354	0.351	0.263	0.241
$Semfeat^{FG}_{Caffe}$	0.461	0.440	0.384	0.354	0.273	0.243
$Semfeat^{IN}_{VGG}$	0.478	0.448	**0.399**	**0.377**	**0.306**	**0.282**
$Semfeat^{FG}_{VGG}$	**0.501**	**0.487**	**0.443**	**0.417**	**0.337**	**0.307**

TABLE 4.3: Results for CBIR runs with the NUS-WIDE dataset. $Semfeat^{IN}$ and $Semfeat^{FG}$ results are reported for sparsification $K = 200$ and query sparsification $K_{query} = 200$.

	Metric: MAP@k					
	k=5	*k=10*	*k=50*	*k=100*	*k=500*	*k=1000*
Caffe	0.133	0.129	0.118	0.115	0.097	0.083
VGG	0.147	0.131	0.124	0.116	0.098	0.086
MC	0.248	0.169	0.150	0.134	0.118	0.095
PiCoDes	0.260	0.173	0.152	0.134	0.116	0.094
$Semfeat^{IN}_{Caffe}$	0.335	0.290	0.237	0.221	0.188	0.173
$Semfeat^{FG}_{Caffe}$	0.344	0.314	0.268	**0.261**	**0.218**	0.192
$Semfeat^{IN}_{VGG}$	**0.351**	**0.317**	**0.274**	0.260	**0.218**	**0.199**
$Semfeat^{FG}_{VGG}$	**0.368**	**0.343**	**0.296**	**0.285**	**0.250**	**0.224**

large margin. The same behavior can be seen in Table 4.1. For the experiments carried on MIRFLICKR and NUS-WIDE, we chose not to repeat the tests with Fisher, Overfeat and $Semfeat_{Overfeat}$ descriptors. However, for these two test collections, we evaluated the retrieval performance other two state-of-the-art semantic descriptors (MC and PiCoDes), detailed in Section 4.3.2. We also test with MAP@k, with k in *5, 10, 50, 100, 500, 1000.*

On MIRFLICKR, VGG descriptors are second best to $Semfeat^{FG}_{VGG}$ for MAP@5 and MAP@10, while MC and PiCoDes fall behind all *Semfeat* configurations. When comparing MC and PiCoDes, we can not draw a clear conclusion. PiCoDes are better on NUS-WIDE, whereas MC are better on MIRFLICKR.

It is important to notice here that harvesting social intelligence under the form of Flickr groups has a direct advantage over manually labeled ImageNet concepts when designing semantic features for any of the three initial image representations (Overfeat, Caffe and VGG). More interesting, these pattern can be observed on all of the three evaluation datasets. We performed a t-test statistical significance evaluation between the following

TABLE 4.4: Best and worst 10 topics ranked by MAP score using $Semfeat_{80}^{FG}$ with *origGT* on the Wikipedia Retrieval 2010 dataset.

	MAP range	Textual topics
Best 10	0.52 - 028	stars and galaxies, tennis player on court, close up of bottles, polar bear, cyclist, race car, launching space shuttle, lightning in the sky, civil airplane, sailboat
Worst 10	0.003 - 0	paintings related to cubism, fractals, musician on stage, DNA helix, shiva painting or sculpture, solar panels, Oktoberfest beer tent, Rorschach black and white, videogame screenshot, Chernobyl disaster ruins.

pairs of features: $Semfeat_{VGG}^{FG}$ and $Semfeat_{VGG}^{IN}$, $Semfeat_{VGG}^{FG}$ and VGG and, finally, between $Semfeat_{VGG}^{FG}$ and MC. We find that: $Semfeat_{VGG}^{FG}$ is significantly different from $Semfeat_{VGG}^{IN}$ with p at least 0.01 on MIRFLICKR and p at least 0.05 on NUS-WIDE; $Semfeat_{VGG}^{FG}$ is significantly different from VGG with p at least 0.001 both on MIRFLICKR and on NUS-WIDE; $Semfeat_{VGG}^{FG}$ is significantly different from MC with p atleast 0.001 both on MIRFLICKR and on NUS-WIDE.

4.6.2 Results analysis

The query set includes 70 topics and it would be therefore impractical to plot individual bars in order to visualize results. Instead, we present best and worst 10 topics ranked by MAP scores in table 4.4. Confirming intuition, topics with high MAPs correspond to topics commonly depicted in Flickr and are thus well represented $Semfeat_{Overfeat}^{FG}$. Topics with low MAPs often depict non-natural scenes and the bad behavior of $Semfeat_{80}^{FG}$ is explained by two factors: (1) *Overfeat* was trained mostly with natural images and (2) non-natural image topics are poorly represented in Flickr groups. CNN retraining would be needed to deal with these cases but it falls outside the scope of the chapter. Other examples of bad behavior include topics which are visually hard. For instance, *Oktoberfest beer tent* images often depict crowds which are difficult to distinguish, *solar panels* are visually similar to the surface of *skyscrapers* while *Chernobyl disaster ruins* can easily be mistaken for other ruins.

To give insight about $Semfeat_{Overfeat}^{FG}$ robustness on the Wikipedia Retrieval 2010 dataset, we compare its individual topic MAPs with those of the two baselines . The first comparison shows that $Semfeat_{Overfeat}^{FG}$ is better in 45 cases (average MAP gain of 0.068), *Fisher* in 22 cases (average MAP loss of -0.031) and there are 3 ties. The largest 3 gains are obtained for *tennis player on court* (0.469), *cyclist*(0.41) and *polar bear* (0.391).

Inversely, the largest performance losses occur for *postage stamps* (-0.177), *brain scan* (-0.099) and *earth from space* (-0.095). These examples confirm those presented in table 4.4 and indicate that $Semfeat_{80}^{FG}$ is better for natural images whereas *Fisher* behaves better for other types of images. The second comparison shows that $Semfeat_{Overfeat}^{FG}$ is better in 40 cases (average gain of 0.037), *Overfeat* in 27 cases (average loss of -0.031) and there are 3 ties. The largest 3 gains are obtained for *lightning in the sky* (0.239), *sharks underwater*(0.232) and *surfing on waves* (0.1785). Inversely, the largest performance losses occur for *notes on music sheet* (-0.1713), *flying hot air balloon* (-0.1582) and *Saturn* (-0.098).

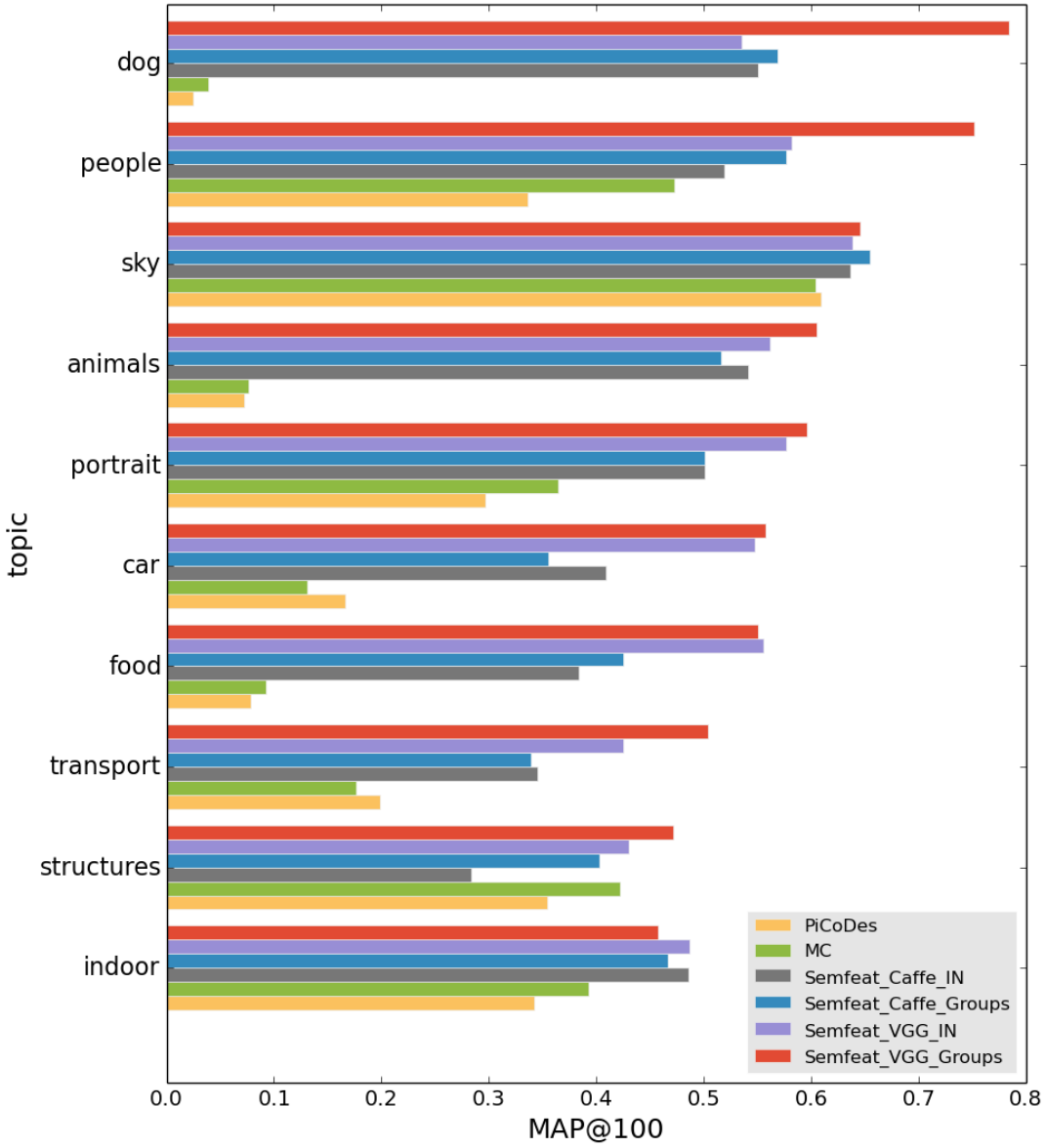


FIGURE 4.9: Best 10 queries on the MIRFLICKR dataset for $Semfeat_{VGG}^{FG}$. We report the results for MAP@100. For each topic, we also report the corresponding results for $Semfeat_{VGG}^{IN}$, $Semfeat_{Caffe}^{FG}$, $Semfeat_{Caffe}^{IN}$, MC and PiCoDes.

In Figure 4.9, we detail the results for the best 10 queries on the MIRFLICKR dataset, while in Figure 4.10, we detail the results for the best 10 queries on the NUS-WIDE dataset for $Semfeat^{FG}_{VGG}$. We report the results for MAP@100. For each topic, we also report the corresponding results for $Semfeat^{IN}_{VGG}$, $Semfeat^{FG}_{Caffe}$, $Semfeat^{IN}_{Caffe}$, MC and PiCoDes.

When analyzing the queries where *Semfeat* has the best MAP scores, both in Figure 4.9 and in Figure 4.10, we notice visually salient diverse concepts. Coincidentally, *sky* is ranked third both for MIRFLICKR and NUS-WIDE. We also notice that the *animal* topic is high in both datasets (first for NUS-WIDE and fourth for MIRFLICKR). One of the main difference between the two evaluation collections is the gap between our proposed semantic descriptors and the other two semantic descriptors (MC and PiCoDes). In Figure 4.9, we note that for several queries (*e.g. people, sky, indoor*), MC and PiCoDes have MAP scores close to those of *Semfeat*. In the case of the *structures* query, MC even surpasses $Semfeat^{FG}_{Caffe}$ and $Semfeat^{IN}_{Caffe}$. On NUS-WIDE (Figure 4.9), however, the difference is constantly much higher between MC or PiCoDes descriptors and any *Semfeat* configuration. For example, for the *animal* topic, MC or PiCoDes have both a MAP@100 under 0.3, whereas all four tested Semfeat configurations achieve a MAP@100 of 1.

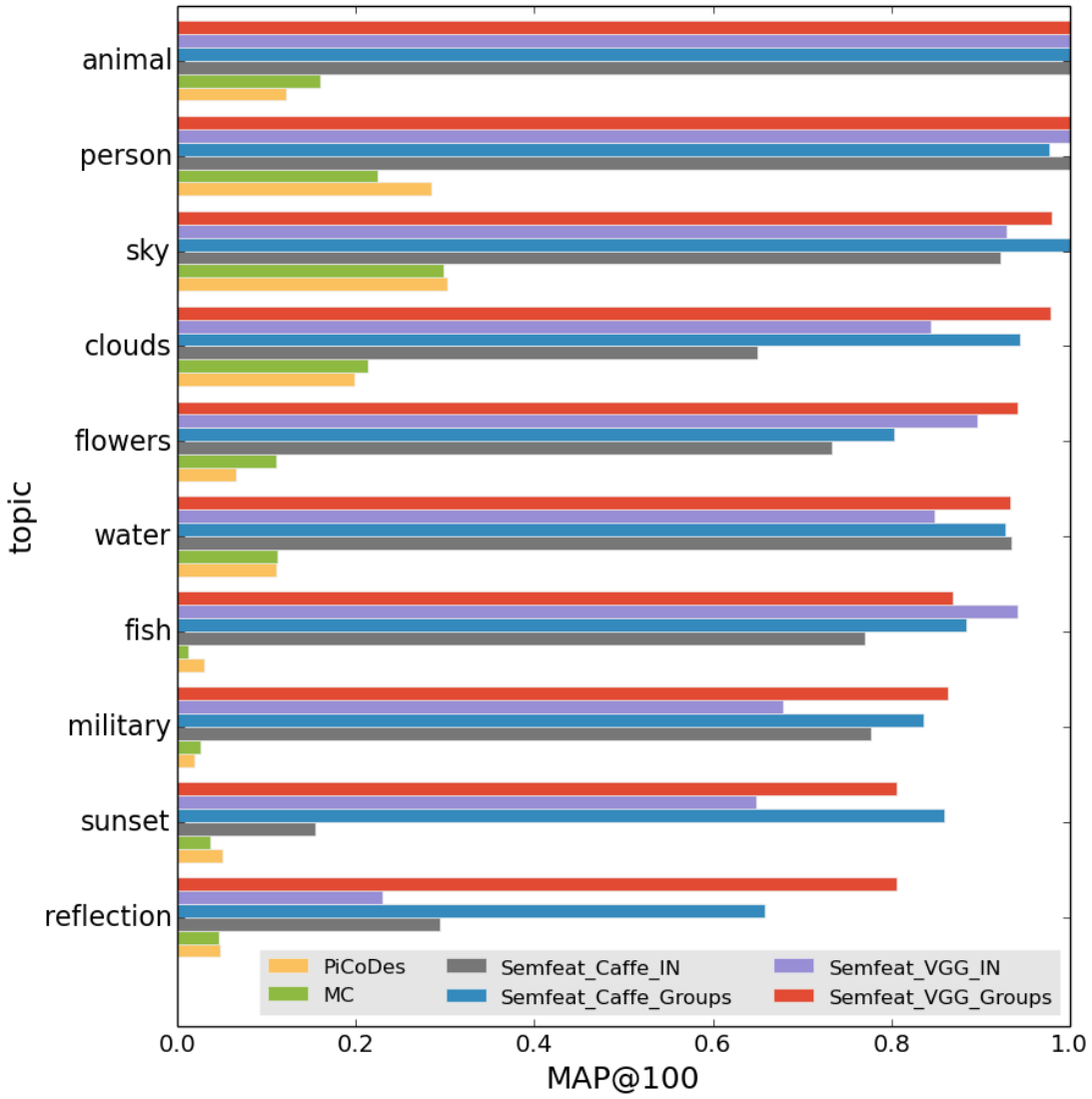


FIGURE 4.10: Best 10 queries on the NUS-WIDE dataset for $Semfeat^{FG}_{VGG}$. We report the results for MAP@100. For each topic, we also report the corresponding results for $Semfeat^{IN}_{VGG}$, $Semfeat^{FG}_{Caffe}$, $Semfeat^{IN}_{Caffe}$, MC and PiCoDes.

We further illustrate the results obtained with $Semfeat^{FG}_{Overfeat}$ on the Wikipedia Retrieval 2010 dataset in figure 4.11. The first two rows have high MAP scores in the original ground truth, the following two are in the middle of the topic ranking and the last two rows correspond to queries with poor results. Although not in focus in this chapter, automatic image annotation with large vocabularies is a part of *Semfeat* pipeline and we present a list of 5 Flickr group tags which are automatically associated to query images. Interestingly, even though annotations are only partially relevant, their combination in *Semfeat* often favors the retrieval of relevant images, as this is the case for *tennis player*. The only image whose annotations are all conceptually unrelated to the image appear

for *DNA helix*. Along with the low MAP examples from table 4.4, the last two rows indicate that the conceptual support of *Semfeat* should be further extended.

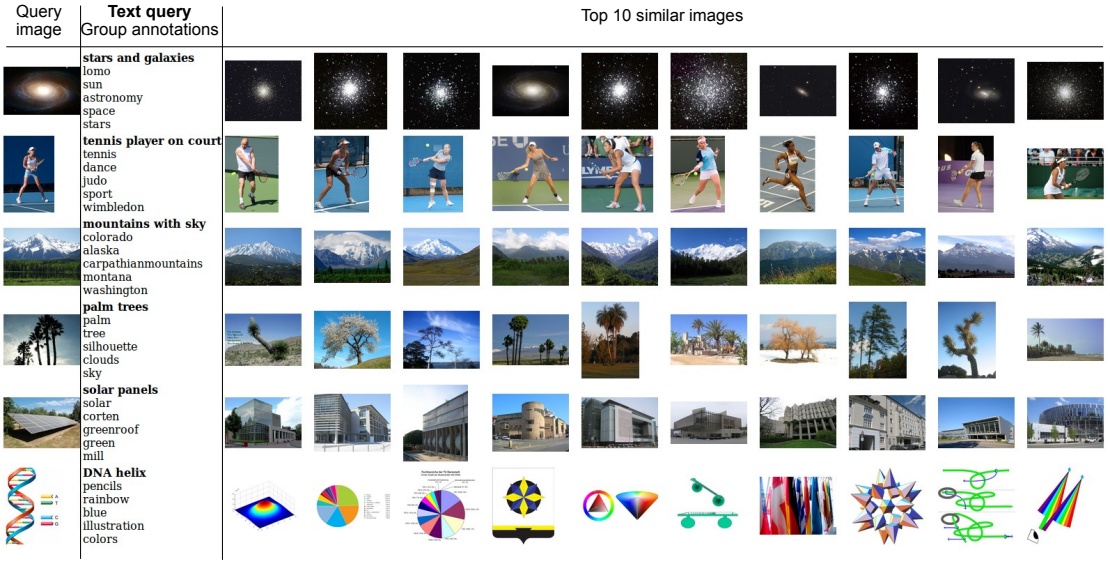


FIGURE 4.11: Illustration of the CBIR process based on $Semfeat^{FG}_{Overfeat}$. We present the query image, the associated textual topic (bold face), 5 automatic annotations from Flickr groups and the most similar images from the Wikipedia collection. We present two highly ranked topics, two from the middle of the ranking and two from the bottom according to *origGT*. The *mountain* example illustrates well the incompleteness of the *origGT* because, while relevant, many of its neighbors were not found by official campaign runs.

4.7 Retrieval Scalability

Previous experiments showed that *Semfeat* obtains high-accuracy retrieval and outperforms previous state-of-art features. In the following, we investigate its behavior in a large-scale retrieval context. Since *Semfeat* is sparse, it can benefit from an inverted file index as described in Subsection 4.2.2. We run scalability experiments for image retrieval using an inverted index on a single core using Intel Xeon CPU E5-2643 @ 3.30GHz processors and a 256GB RAM. A naive in-memory C++ implementation of the inverted index is used in which inverted index concepts are stored as keys and associated images, with scores, are stored as values.

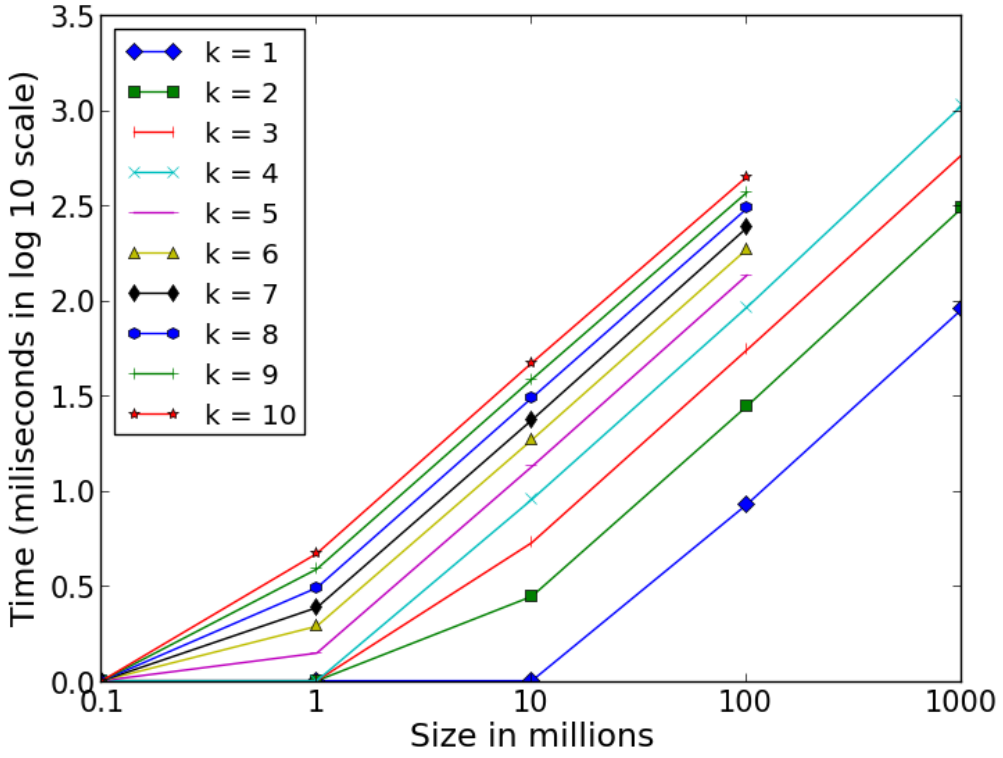


FIGURE 4.12: Search latency with sparsification $K \in 1, ..., 10$ and simulated dataset sizes up to 1 billion images. To improve visualization, log_{10} scaling of latency (in milliseconds) is used. Values are averaged over 1000 query images.

To evaluate retrieval performance, large datasets are simulated by duplicating several times the Wikipedia Retrieval images collection, an operation that has no incidence on retrieval speed. The duplications result in the image set $\{{I_1}^1, ..., {I_1}^d, ..., {I_M}^1, ..., {I_M}^d\}$, where I_i is the i-th image in the original dataset, M is the original dataset size ($M = 10^5$ for Wikipedia Retrieval) and d is the dataset duplication number. *Semfeats* representation are extracted from the images and represented using an inverted index. Due to sparsity, m_c, the number of images that are associated to the c-th concept in the inverted list, respects $m_c << M$. For $M = 10^5$ images and $d = 1$, the average number of images per inverted list is 61.84, with a standard deviation of 146.33 and a maximum value of 3402. In our experiments, we set $d \in 1, 10, 100, 1000, 10000$ to investigate retrieval search latencies for collections including between 100,000 and 1 billion images.

A set of 1000 image queries is randomly selected from the Wikipedia collection and its images are retrieved against the different simulated collections. To assess the sparsity on the representation, we repeat this experimentation for *Semfeat* with different sparsification factors ($K \in [1, 2, 3, 4]$). Inverted indexes are more compact for small values of K but the retrieval performance is equally more reduced (see Section 4.4). The average search latency (in milliseconds) in the inverted file is measured for each d and K

combination. Results are reported in figure 4 using a log_{10} scaling. For 1 billion images, retrieval latencies are calculated for *Semfeat* using $K \in [1, 2, 3, 4]$ due to RAM limitation. This limitation is relatively easy to circumvent through index distribution or pruning but these operations are beyond the immediate scope of the chapter.

When the collection size is up to 10 million images, real-time searches are supported for every value of K. For 10 million images, average latency is 0.85, 13.6 and 47.8 ms for $K \in \{1, 5, 10\}$. Corresponding latencies for a collection of 100 million images are 8.6, 130.8 and 454 ms. For comparison, a brute force search with *Overfeat* in simulated datasets of 10 and 100 million images is approximately 10 and 100 s, a much larger latency for an equivalent accuracy. This result shows that *Semfeat* enables near real-time searches in a 1 billion images collection using a single core, provided that enough RAM is available. In our implementation, RAM consumption for 1 billion images is between 62 GB ($K = 1$) and 248 GB ($K = 4$). The authors of [25] report results for a 128 bit compression of a Fisher Kernels using a 100 million images collection. It is not possible to directly compare the accuracy of both approaches but we can compare latencies. In [25], average search latency is 250 ms while in our approach it ranges between 8.6 and 454 ms, depending of the chosen value of K.

4.8 Discussion

4.8.1 Advantages

CBIR results show that the *Semfeat* versions based on reranked Flickr groups outperform other existing methods which were tried on the Wikipedia Retrieval dataset. To our knowledge, this is the first time when a CBIR result is not lagging far behind text approaches tried on the a complex collection (textual run $MAP = 0.2361$ [24] vs. CBIR $MAP = 0.2127$ here).

Beside competitive performances and contrary to widely used image features, such as bags-of-visual-words, Fisher Kernels [25] or CNN features [14], *Semfeat* directly conveys semantic meaning. Image similarities are based on the comparison of humanly understandable dimensions (i.e. Flickr groups or ImageNet concepts), a characteristic which enables result explainability. Given a query and a result, users can browse the list of common concepts in order to have the semantic aspect of the image, thereby participate to reduce the *semantic gap*.

Another advantage of *Semfeat* is its sparsity. The best performances are obtained when only a few dozens of concepts are kept for each image. In this configuration it is straightforward to efficiently represent images as inverted indexes in order to speed up retrieval. We tested inverted search with a simple in-memory C++ implementation and simulated datasets up to 100 million images with sparsity $K = 100$. Retrieval time grows linearly and is under 1 millisecond for 10 million images and under 10 milliseconds for 100 millions. For comparison, we also tested forward search with Overfeat (4096 dimensions) and obtained a retrieval time in the range of 15 seconds for 10 million images. Even if one would use compressed versions of dense features, inverted search would still be faster.

Last but not the least, $Semfeat^{FG}$ is built on top of an automatically mined dataset. We deliberately use simple but efficient techniques to rerank images and learn models. The proposed pipeline facilitates resource extension, with the only limitation being the availability of sufficiently large image sets for new groups or concepts.

4.8.2 Limitations

We have mentioned some *Semfeat* limitations in the experimental section and extend the analysis. The learning methodology used here is scalable but can be improved. With the use of more sophisticated models, the predictions associated to *Semfeat* dimensions would probably be more robust and have a positive impact on the overall results. However, when choosing the learning models to use, one should keep in mind that the prediction process needs to be fast, a constraint which is particularly relevant when the semantic features include a large number of dimensions. Consequently, with the authors of [23], we advocate for the use of linear models.

Another limitation is the choice of positive examples which model individual groups/-concepts. We implemented a first version of *Semfeat* with a maximum of 300 images per groups. While this volume is sufficient for simple visual concepts, it is probably insufficient to model complex concepts and future experiments should focus on enlarging the positive examples set. It is particularly interesting to investigate whether reranking methods can benefit from a larger number of available examples. In such a setting, a larger amount of potentially noisy images could be removed while still having a sufficiently rich and diversified representation of the concept.

Flickr groups, which are created in an unsupervised manner and they mirror users' interests but are often redundant. For instance, there are tens of different groups which focus on *classic cars* and several of them can be jointly activated in the *Semfeat* representation of the same image. These groups could probably be merged into larger meta-groups in

order to reduce redundancy and propose more informative features. Redundancy reduction is also applicable to the combination *Semfeat* features obtained with different resources, including ImageNet and Flickr groups. Similarly to the diversification problem in image retrieval, a fine balance needs to be found between precision and diversity, two characteristics which are often contradictory. This non-trivial problem was not tackled in this Thesis and is left for future work.

4.9 Conclusion

4.9.1 Contributions

In this Chapter, we proposed a technique for the automatic mining of large-scale visual resources from Web data. Based on this result, we proposed a new semantic image representation which was tested in a content based image image retrieval task. Returning to our initial research questions we can conclude that:

Q1 With an appropriate choice of the initial collection and with the introduction of scalable but efficient image reranking techniques, the results obtained with the automatically built resource can rival with those of the manual resource. While it needs confirmation in other image mining tasks, this finding has important implications for the way visual resources are built and exploited. At large scale, automatic resource construction requires significantly less effort than manual labeling and constitutes an appealing alternative to datasets such as ImageNet.

Q2 Efficient semantic representations can be built through the combined use of powerful initial features, such as CNNs, and of an appropriate visual representation of feature components. Further investigation is needed concerning the choice of machine learning models and the number of images, both positive and negative, used for learning individual models. It is probable that more sophisticated models combined with a larger number of training images will improve results. A good coverage of the conceptual space can be obtained through an appropriate choice of the initial Web dataset. We have tested the use of Flickr groups but other large concept sets, such as Wikipedia article titles, could be considered. The pipeline presented here is easily applicable to larger amount of Web images, the only potential constraints being related to the availability of data and to the processing power needed to build individual models. With the use of simple scalable learning models, as proposed here, it is easy to scale way beyond tens of thousands of models. However, when enriching conceptual coverage, one should be careful about the potential negative effects of redundancy, a problem which deserves close attention.

Q3 In image retrieval, compactness is achieved by sparsifying semantic features and by using inverted indexes. With this scheme, very large volumes of data can be searched without precision loss, as it is the case for existing compact features [25].

Q4 The results reported here indicate that semantic features are very useful for CBIR retrieval. Following [22], [23] or [3], we bring new evidence concerning the usefulness of semantic features and, in particular, propose an efficient way to clean and exploit large-scale noisy Web corpora.

4.9.2 Perspectives

In this Chapter, we investigated the use of semantic features for CBIR. As a complement to CBIR, it is natural to evaluate *Semfeat* performances in an image classification task. In a preliminary experiment, we used the publicly available Pascal VOC 2007 dataset [353]. It includes 20 object classes whose instances are collected from Flickr and manually annotated. For each class, a linear classifier is trained by stochastic gradient descent [354] with a one-versus-all strategy. The coefficient C of the data fitting term was fixed to 10^{-4} based on validation on an independent database. Performances are evaluated with mean average precision (mAP).

TABLE 4.5: MAP classification performances on PascalVOC 2007. After preliminary *Semfeat* features are used with sparsification $K = 100$.

Overfeat	0.711
$Semfeat^{FG}_{Overfeat}$	0.718
$Semfeat^{IN}_{Overfeat}$	0.736

Noticeably in our experiments the *Overfeat* network was trained with the ImageNet ILSVRC 2013 database without adaption to the VOC Pascal dataset as recently proposed in [318]. These authors report a $MAP = 0.777$ and thus validate the usefulness of adaptation. We instead use semantic features in order to improve visual recognition. We compare classification performances obtained by the two versions of *Semfeat* against *Overfeat*. The results reported in table 5.7 show that only a slight improvement is obtained compared to *Overfeat*[5]. Contrarily to the CBIR task, $Semfeat^{IN}_{Overfeat}$ obtains the best performances for classification. This behavior is probably explained by the fact that PascalVOC concepts are mapped better in ImageNet than in Flickr groups. The follow-up this experiment with *Semfeat* descriptors built from models based on Caffe

[5]All results are reported with the small net *Overfeat*. The use of the large network would slow down feature extraction but would also increase performances [14].

and VGG features, together with testing on other classification datasets is left for future work.

Another idea which we did not test in this Chapter is the adaptation of the negative sets to the target positive concept, similar to the approach presented in [355]. The learning process would benefit from the use of negative images representations which are close to the classifier margin. However, negatives should be carefully chosen in order for them not to be in a hierarchical relation with the positives. For instance, it would be counter-productive to learn a model for *dog* with any dog species as negative example.

Finally, we focused uniquely on the exploitation of visual image characteristics and did not address the combination of textual and visual cues. If done appropriately, this combination can have beneficial results in image mining [24] and will be tested in the future.

Chapter 5

User credibility in image sharing platforms

In this Chapter, we first define the concept of user tagging credibility in the context of Flickr users. Another contribution of this part is the creation of a new dataset specifically built i) to help us evaluate potential indicators for credibility and ii) to serve as a training dataset on which one can compare multiple learning models and features. We also detail important statistics on the number of users, images and rater agreement scores. Our main goal is to propose multiple features that can serve as estimators for user credibility. We extract both context and content features stemming from various data sources (Flickr groups, photo favorites or a user's contacts network) and discuss a total of 66 credibility estimators. We also exploit the set of concept classifiers introduced in Chapter 3 to account for the relations between image tags and its visual content. We furthermore describe the data acquisition process for a large set of features and test their usefulness as individual credibility estimators. Finally, we define a credibility prediction problem, in which we learn regression models that provide better credibility estimators than the individual features.

5.1 Motivation

While the analysis of users of social media websites, such as Twitter [231, 243, 258–261], Facebook [254] or blogs [146] is well studied, there are few works that directly target users in image sharing platforms. In this Chapter, we approach the study of

user tagging credibility in this setting. Since its early years, Flickr has been widely used as a playground for identifying the motivations and goals of users for tagging their images. Although we cannot directly exploit the findings reported in these types of studies directly, they offer both important insights on how we can link previously defined user categories to credibility, as well as a theoretical motivation for exploring different data sources when proposing context features.

In a seminal work, Ames and Naaman [356] propose a taxonomy of motivations for annotation along two dimensions (sociality and function), and explore the various factors that people consider when tagging their photos. They base their work on user interviews and other qualitative methods. The first dimension, *sociality*, relates to whether the tag's intended usage is by the individual who took and uploaded the photo or by others, including friends/family and strangers. The second dimension, *function* refers to a tag's intended uses. They found that users tagged their pictures either to facilitate later organization and retrieval or to communicate some additional context to viewers of the image (whether themselves or others). The *function* dimension focuses on the motivation for adding tags.

In this Chapter, we look for user credibility indicators that can improve an image retrieval framework. This is why, when considering the *sociality* dimension, we are interested in discriminating between users that upload their images to Flickr with the goal of showing their contributions to the community and those that do it to have a backup of their photo collection or only show them to a limited set of people. Contributions from the first category of users are potentially more useful to be used for general purpose retrieval than those coming from users falling under the second category. Similarly, when looking at the *function* dimension, we are interested in identifying users that have as their main purpose for tagging the improvement of retrieval over their photo collection. In contrast to this type of users, we would like to be able to filter out the users that tag mostly to indicate the context in which the photo was taken (e.g. the year when the photo was taken or the names of people attending a certain event) which would be relevant only to their immediate social circle. In [357], the authors also investigate the motivations behind user tagging and propose several incentives that can be used in an annotation framework. They argue that annotators gain widespread recognition and credibility by doing good work that can be used by many people, even if they are not receiving direct compensation for their work. In our work, we use the concept of user tagging credibility in a similar way. We consider a user to be credible if his/her contributions are useful for the community. Previous works investigating the motivation behind tagging offer clues on which aspects of the Flickr framework to investigate when looking to extract credibility estimates. For example, the community recognition argument mentioned by

Kustanowitz and Shneiderman [357] can be found in a user's desire to have his images included in Flickr groups.

5.2 Problem description

We consider each of credibility components introduced in Chapter 2 to be domain specific. Depending on each case, we may define credibility through one or several of these components. In this Thesis, we focus on user credibility in image sharing platforms and, in this context, user credibility is mainly reflected in the quality of a user's contributions. Following the model described in Figure 2.1, we identify each credibility component for user image tagging credibility (in particular to the Flickr platform) as follows:

- **trust:** refers to how a user is perceived by the community. Indicators of trust may include the number of users that have him or her among their contacts, or the comments the user receives for his/her photos.
- **expertise:** either real life photography expertise or validation received by the community. Indicators of expertise may include clues in the user's description (*e.g.* working for a professional photography institution) or being invited to exclusive curated Flickr groups.
- **quality:** for the credibility facet approached in this Thesis, we examine the quality of image tagging and not the photos themselves. Imposing this restriction for the term *quality*, we consider an image to have good quality tags if they well correlated with the visual content of the image. We note here the difference from *truthfulness*. For example, a user may tag his images with the type of camera they were taken with or the date when the photos were taken. While relevant for the user, these tags serve no purpose for describing the content of the image and cannot be used in a retrieval scenario.
- **reliability:** The sustained tagging quality (following the definition presented in the previous item) of a user's images in time.

We share the observation made by Ye and Nov [358] who state that researchers need to take a user-centric approach to understanding the dynamics of content contribution in social computing environments. They study the connection between different aspects of a user's motivation and both the quantity and quality of his contributions. Their results indicate, among others, that users with more social ties, especially ties with people they have not met in the physical world, tend to contribute better content to the community.

In order to estimate a user's credibility score, we investigate two complementary cues. We explore both content (*e.g.* tags and images) and context features (*e.g.* Flickr groups, photo favorites or a user's contacts network).

5.3 A Multi-Topic Tagging Credibility Dataset (MTTCred)

We propose in this section a novel dataset, designed with the goal of analyzing user credibility for a diversified set of topics.

5.3.1 The need for a dedicated user tagging credibility dataset

Having an indication on the credibility of the source can be beneficial for the performance of an image retrieval system. This has been recently proven by the introduction of user credibility in the 2014 MediaEval Retrieving Diverse Social Images Benchmarking Initiative [20], where some of the participating teams [288, 289] have improved the relevance and diversity of an image retrieval system using user credibility estimators. This Benchmarking Initiative also offers the only available dataset that provides manual credibility estimations for Flickr users, the Div150Cred dataset [290]. It provides Flickr photo information (the date the photo was taken, tags, user's id and photo title , the number of times the photo has been displayed, url link of the photo location, GPS coordinates) for about around 300 locations and 685 different users. Each user is assigned a manual credibility score which is determined as the average relevance score of all the user's photos. To obtain these scores, 50 157 manual annotations are used (on average 73 photos per user). Although the aforementioned works are groundbreaking in their use of user credibility estimates for image retrieval, this process is only performed in a confined setting (*i.e.* diverse image search for the tourist domain.

In this Thesis, we go beyond the direct usage of user credibility estimators in an image retrieval system and propose a medium for a complex analysis of user tagging credibility that can serve multiple purposes, including credibility class prediction and credible user ranking. We provide ground truth credibility estimations for 1009 users whose evaluated images cover a large set of visually coherent topics. Our proposal diverges from recent image retrieval datasets that are either domain specific [359] or built for ad-hoc retrieval of complex topics [360]. It is closer in terms of topic coverage to the original MIR Flickr collection [361]. Our target is to obtain a reliable collection of ground truth scores for user credibility and not proposing another image retrieval dataset.

Throughout this chapter we use Div150Cred to provide provide comparative feature analysis. For a detailed description of this dataset, we refer readers to [290]. Such a comparison is useful to assess the reliability of features across domains.

5.3.2 User credibility dataset design

We describe here the main requirements for creating a dataset tailored for the investigation of features that are potentially useful in assessing a user's tagging credibility:

- It should contain contributions from a substantial number of different users. This allows the exploitation of the dataset both as a relevant collection on which correlations between automatically extracted features and manual credibility scores can be estimated, but also leaves room for a learning scenario in which the credibility score can be predicted by a trained model. It should offer enough training instances so that commonly used machine learning models are able to learn a pattern, if one would exist.
- Each user should have a significant number of contributions evaluated so that we can derive a reliable manual credibility score. This score will be obtained by averaging the relevance scores of individual contributions. This modeling of the manual credibility scores was done to ensure comparability with Div150Cred [290].
- Contributions sampled for each user should be images depicting a diverse set of topics. This choice is imposed by the nature of how we define the credibility score in an image tagging context, i.e. as a global property of a user's contribution. Having more than one topic represented for each user also promotes the re-usability of this dataset and enables studies on domain specific user credibility. We do not investigate domain specific credibility in this Thesis but this dataset will allow this research direction to be explored in future work.

In practice, all of the desired features mentioned above are subject to limitations coming from the availability of data but mostly from the cost of annotation. As a result, when setting the targeted values for each of the three features, a trade-off between any of them has to me made. After a series of internal studies, we settled for the following approximate values: approximately 1000 users, 50 images for each user and at least 5 topics represented in the contributions evaluated for each user. Next, we present the dataset annotation protocol and the dataset statistics.

5.3.3 Dataset creation

We follow an annotation methodology similar to that proposed for the construction of the datasets used in the ImageCLEF Wikipedia retrieval evaluation campaigns [360]. For each topic, we present the annotator with a couple of relevant images and a narrative which has the purpose of clarifying what is relevant and what is not for each topic. For example, in the case of the *sun* topic, we provide the following narrative: *Assume that you want to illustrate different aspects of sun with images. Please select all images which are relevant for sun from the list below. Diversified views or aspects of sun are relevant.* In Figure 5.1, we also show an example of how topic relevant images are presented to the annotators.

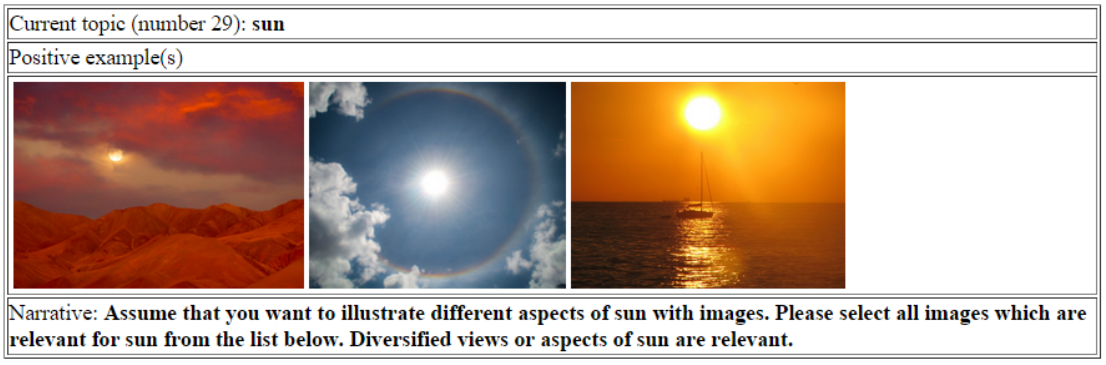


FIGURE 5.1: Example of the sample images and narrative given to the annotators.

Then, for each topic we present a maximum of 300 images on a single page whose interface is similar to that of search engines and offers the possibility to easily select relevant images. The relevance assessments of the images in the dataset were provided by a total of 6 trusted annotators (faculty members), with 3 annotations per image. An image is considered to be relevant if at least two raters agree upon it. Before starting the annotation process, the users were first involved in a feedback loop. This entailed them expressing the ambiguities they identified in some topics and, from our side, modifying the narratives, where necessary. We first fix a number of diverse but simple topics that have a clear visual representation, that are illustrated in Figure 5.2. This means having confident assessments of images depicting easily recognizable topics.

We use the Flickr API[1], to download both user and image metadata. We start with the *flickr.photos.search* function to download photo metadata (limited to Creative Commons) for approximately 90 topics. Then, we collect statistics on the users that have contributions to the retrieved set of images for all the topics. We retain the users with most images across topics. We keep the top 3000 users as candidates for the credibility

[1]http://www.flickr.com/services/api/

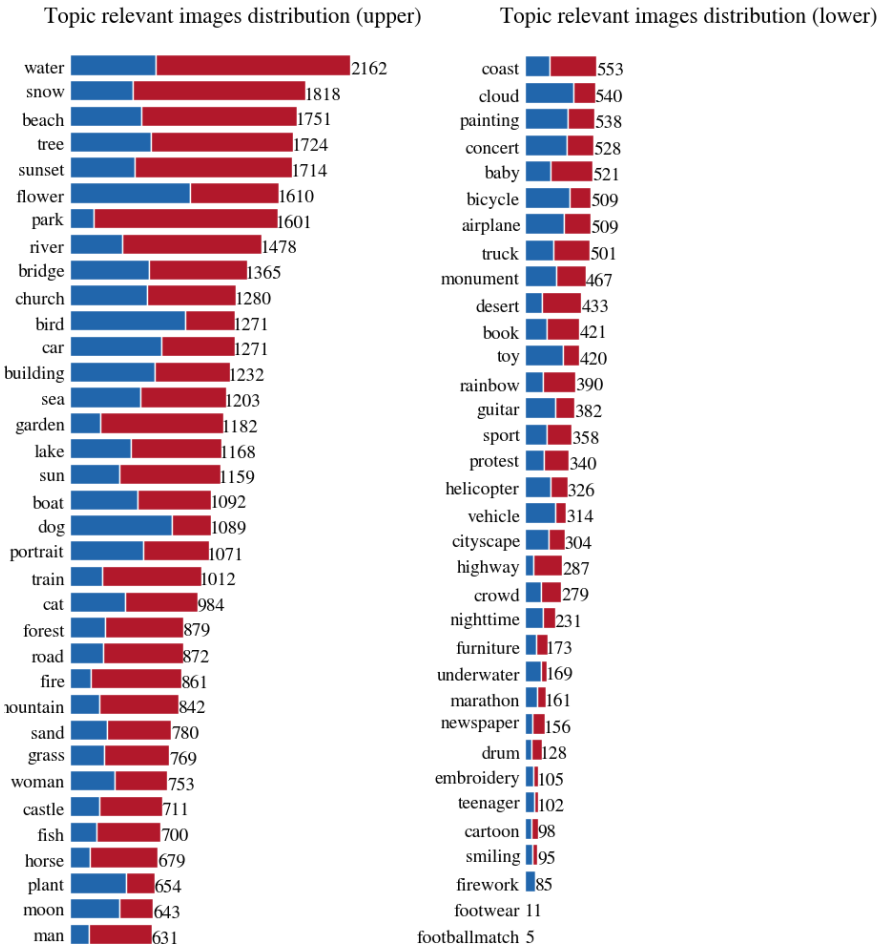


FIGURE 5.2: Distribution of relevant and non relevant images for each topic

dataset. For each of these users, we call the *flickr.people.getPhotos* function to gather metadata for the users' photos. We download metadata for a maximum 10 000 images per user. Finally, we keep only the users that have at least 50 images covering at least 10 topics.

5.3.4 Dataset statistics

Using the protocol described above, we obtain a dataset containing a total of 1 009 users and 50 450 images evaluated for relevance covering 69 topics. The remaining topics were covered only by few users and have been discarded. Each user has exactly 50 images in the dataset that will be manually evaluated. In Figure 5.2 we present the names of the assessed topics and the number of images that were evaluated for each topic. We also show the distribution of positive and negative images for each topic. The blue bar represents the proportion of images found relevant by the annotators and the red bar

gives the proportion of non-relevant images. Following general trends in Flickr[2], some topics are very frequent (e.g. *water, snow, beach*), while others have fewer than 100 images (e.g. *firework, footwear, footballmatch*). We can also see from this figure that a majority of the images are rated as being non-relevant to the tags. A few notable exceptions, where the relevant images are predominant are the *dog, plant, vehicle or firework* topics.

We observe the agreement between raters by measuring Randolph's free marginal multirater kappa score [362]. We use this method to evaluate agreement, as opposed to Fleiss' multirater kappa, because we do not know a priori the quantities of cases that should be distributed into each category (relevant vs. non relevant images). We observe an agreement score of 0.581 when combining annotation for all the topics, which can be interpreted as moderate to high agreement [363]. This score shows that although we took precautions to ensure a simple and clear annotation process, providing relevance ratings for a diverse set of topics remains a difficult task.

TABLE 5.1: Randolph's free marginal multirater kappa score for individual topics.

Topics with high agreement			**Topics with low agreement**		
Name	***Kappa***	***#Images***	***Name***	***Kappa***	***#Images***
fire	0.86	861	truck	0.337	501
man	0.854	631	teenager	0.359	102
cat	0.838	984	lake	0.375	1168
marathon	0.776	161	embroidery	0.377	105
rainbow	0.770	390	sea	0.379	1203
helicopter	0.762	326	building	0.393	1232
horse	0.756	679	boat	0.406	1092
vehicle	0.749	314	nighttime	0.411	231
castle	0.735	711	church	0.413	1280
baby	0.733	521	grass	0.422	769

In Table 5.1, we show the first 10 (left column) and last 10 (right column) topics ranked by Randolph's free marginal multirater kappa score for the relevance annotation of the images found in the topic.

As expected, we notice high scores for some of the least ambiguous topics (e.g. *fire, man, cat*). Among the topics with low agreement scores, we find those that may present with some level of incertitude, such as *teenager* but also, surprisingly, topics that seem to have a clear visual representation, such as *boat or truck*.

[2]https://www.flickr.com/photos/tags/

5.3.5 Deriving a ground truth credibility score

Following the process implemented for the *Div150Cred* dataset [290], we compute the manual user credibility scores by taking the percentage of images found relevant among the 50 images that were evaluated for each user. We use 50 images for estimating the credibility score to ensure comparability with the *Div150Cred* dataset. In Figure 5.3 we present the distribution of the manual credibility scores. We observe that the scores follow an approximate normal distribution. The fact that the majority of images are labeled as non-relevant can also be observed in this figure, with a mean of the credibility scores at 0.41.

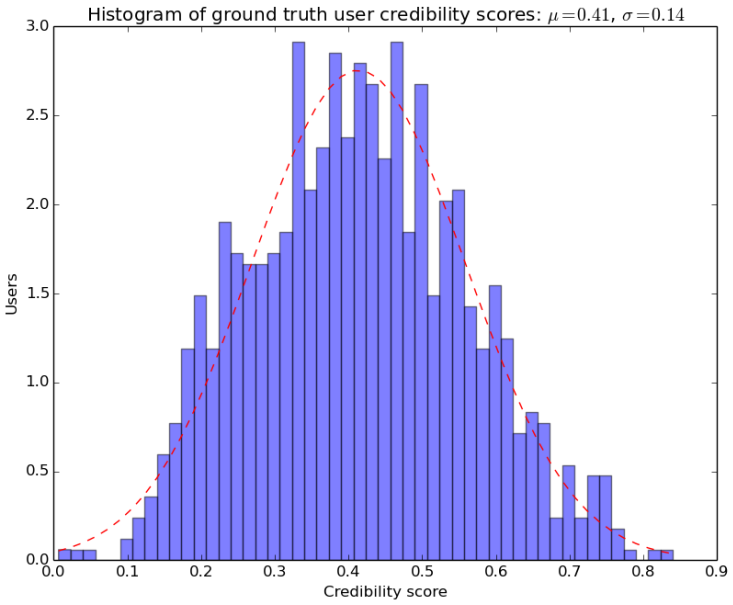


FIGURE 5.3: Histogram of manual credibility scores

In the following, we investigate the impact of having less images on the credibility feature of a user. For this, we simulate a scenario in which we get the credibility estimate by averaging the relevance scores of only k images, with k ranging from 5 to 50. For each user, we randomly pool k images from the full set of 50 annotated images. Then, we analyze the Spearman correlation between these credibility scores and the ones associated to the full set of user images. In order to avoid a bias in the random selection, for each k, we replicate the pooling 100 times.

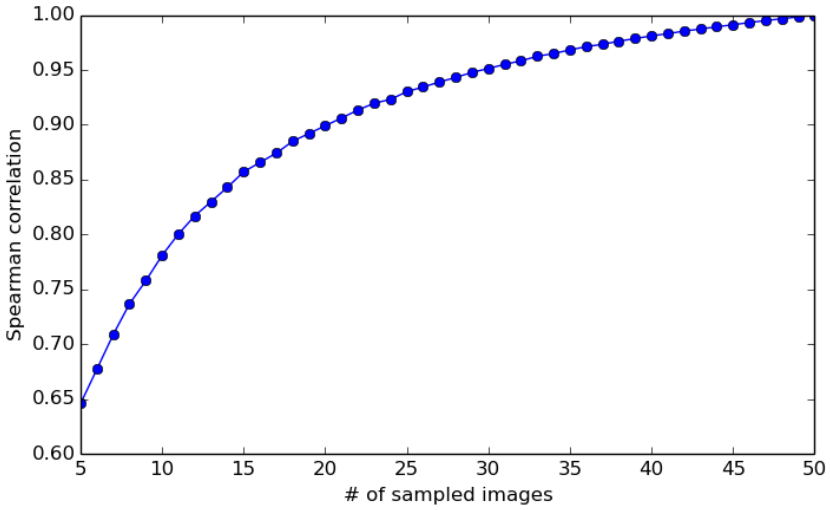


FIGURE 5.4: Spearman correlation between the manual credibility score and a credibility score obtained from subsets of ground truth images of different sizes

In Figure 5.4, we report the average Spearman correlation scores for each k. We note that using only half the images, we get a credibility score that is highly correlated (over 0.9 Spearman correlation value) with the credibility estimates that we get from the full set of annotated images. However, we can observe a slower but continuous increase after including 30 out of the 50 images. This suggests, as expected, that annotating more images for relevance leads to more reliable credibility estimates. However, the slower increase also indicates that selecting 50 images per user gives an acceptable approximation of the credibility of his tags.

5.4 Context features as credibility estimators

Alongside its main functionality of photo storage and sharing, Flickr provides its users with means to organize their photo collection but also to interact among themselves. Besides tagging, users can group their photos in photosets and can add their photos to groups that receive contributions from different users with common interested by the topic. Following a social-network modus operandi, users can have contacts and are able to provide feedback for the photos of other members of the community, through the use of favorites and comments. We are interested to exploit as much of this data as possible for identifying user features that may be good indicators for credibility. Features can be grouped in feature families, according to the nature of the data from which they are extracted. Due to limitations imposed by the number of calls per day to Flickr APIs, we settle for the following feature families: *photo metadata, groups, photosets, given photo*

favorites, and contacts. All the experiments and analysis presented in this section can be easily extended to include features coming from any other data source, when they becomes available. Note that the features we extract can be explicit and may come from a user's direct actions, such as adding a new contact or implicit, which we derive from a user's activity, such as the temporal aspect of his upload behavior.

5.4.1 Data acquisition

Flickr exposes a number of API functions that provide access to a user's contributions or his interactions with other members. Due to API usage constraints, we limit the number of samples we download. This limit is specific to each feature family and is chosen so that we can obtain a significant sample of data for each user. We will now provide details about how we downloaded the collection from which we extract our proposed context features and give statistics on the number of metadata items for each feature family:

- *Photo metadata:* We use the *flickr.people.getPublicPhotos* function of the Flickr API to download metadata associated with a user's photos. For each photo we get data such as the title, tags, the time the photo was taken, and the upload time. We first make a request to retrieve the first page, from which we extract the total number of photos the user uploaded on Flickr. Then, we request 500 items per page for each API call and, if available, we download up to 20 pages for a user. Overall, we collected 10 540 metadata files for the 1009 users in our evaluation dataset, with an average of over 5 000 pieces per user.

- *Groups:* We use the *flickr.people.getPublicGroups* function to get the list of public groups a user is a member of[3]. For each group, we retrieve its name, the number of users that are part of the group and the total number of photos that have been included in the group. For each user, we generate a single file with the data about the public groups he or she is a part of, gathering a total of 1009 group metadata files.

- *Photosets:* Photosets metadata is retrieved by calling the *flickr.photosets.getList*. Similar to photo metadata, we first get the first page, with the number of total photosets and then download a maximum of 10 pages, where each page contains data about at most 500 photosets. This leads to a total of 2 733 downloaded photosets metadata files.

[3]Note that we do not have access to a user's private groups and the collection is limited to public groups.

- *Given Photo Favorites:* We retrieve the list of photos a user has marked as favorite by calling the *flickr.favorites.getPublicList* API function. The procedure is similar to that used for photosets. We obtain 6 337 files, each containing details about up to 500 photos the user had marked as favorite.

- *Contacts:* In Flickr, if a user has added another user as a contact, this action does not imply that there will be reciprocity. The contact relationship is not symmetric. If user A designates user B as a contact, user A can see the photo stream of user B, but not vice versa. This makes the contact relationship closely related to the follower structure in Twitter. The Flickr API provides access only to the contacts of a user but we can not retrieve a list of users that have the target user among their contacts. In order to be able to apply network analysis methods, we crawl a subsample of the Flickr contacts network by recursively calling the Flickr API function *flickr.contacts.getPublicList*. Our methodology for sampling the network is similar to that of Mislove et al. [364]: we start from the list of users in our dataset and we download the contacts up to a depth of 2 (i.e. contacts of a contact of the original user). In order to have a sample of contacts for all of our evaluation users in a reasonable amount of time, we impose a limit to API calls for second degree contacts. If a contact of the original user has more than 500 contacts, we retain only a sample of them. We download the contacts information only for this sample. Using this approach, we obtain a contacts network comprised of 5 811 652 unique users and 91 205 141 links. To put these numbers in perspective, Cha et al. [365] estimate that a network including 2.5 million Flickr users and 33 million links, represents 25% of the entire Flickr network. This statement is made for the Flickr network as of the end of 2007. Newer data suggest that in 2014 there were around 92 million active users in Flickr[4].

Next, for each feature family, we describe the list of individual features that were extracted and motivate their selection. This list is not exhaustive and other features can be easily added from the downloaded collection of data files presented above.

5.4.2 Feature extraction

In this Section we focus only on features that can be extracted from the context. In the case of a Flickr user, we consider the context built around his or her activity to encompass any action he or she performed (except the act of tagging) and any action that concerns him or her done by other users.

[4]http://www.thesocialmediahat.com/active-users

5.4.2.1 Metadata features

The metadata that accompany photos a user uploads to Flickr represent the main source of information about a user's direct contributions in Flickr. Through its API, Flicker provides for each public photo the associated tags, the title given by the uploader, the date it was uploaded, and, if available, the date when the photo was taken. We exploit all of these and extract the following features:

Title related features. Users may choose a different title for one or small number of their photos, or may use the same title for a large set of photos (e.g. all of the photos taken in the same trip). We hypothesize that a user who takes his or her time to provide a detailed title for as many photos as possible, is more likely to provide contributions that are meant to be shared with the community. In the opposite case, when a user attributes few titles for most of his or her photos (i.e. usage of bulk titles), we may be facing a user that only wants to store his or her photo collection, mostly for personal usage. Besides bulk titles, we also investigate the diversity of the vocabulary used in titles. A large diversity of title words may indicate a user who has either interest in a large number of topics or takes his or her photos in many different scenarios. Finally, we look at capitalized words found in titles. A high percentage of capitalized words may indicate a focus on locations or people. We extract the following tag related features:

- *title_bulk_percentage:* the percentage of titles that appear at least 3 times in the set of titles of a user.
- *title_vocabulary_size:* the number of unique words used in the titles given by a user.
- *title_capitalized_words_percentage:* the percentage of capitalized words found in a user's title vocabulary.

Temporal features. A photo upload behavior uniformly distributed over time may be an indicator for a user's constant involvement in Flickr. Temporal data could contribute to separating casual users, who upload images occasionally from those who are more passionate about photography or are professionals. We propose the following time related features:

- *different_upload_days:* the number of unique days in which a user has uploaded at least one photo.
- *average_upload_time_delay_minutes:* the average time elapsed between two consecutive uploads, measured in minutes.

- *average_upload_time_delay_days:* the average time elapsed between two consecutive uploads, measured in days. We look only at the number of days passed between the last upload of one day and the first upload of the next day in which an upload was made.
- *different_photo_taken_days:* the number of unique days in which a user has taken photos.
- *average_photo_taken_time_delay_minutes:* the average time elapsed between two consecutive photo taken timestamps, measured in minutes.
- *average_photo_taken_time_delay_days:* the average time elapsed between two consecutive photo taken timestamps, measured in days.
- *average_date_taken_upload_delay_hours:* the average time elapsed between the time a photo was taken and the time it was uploaded, measured in hours.

For time based features, we wanted to see if there is a difference in the granularity of the scale used to measure the time passed between contributions. To this end, use test both the delay in days and minutes.

Photo related features. Ye and Nov [358] find that in Flickr, the quantity of a user's contributions is negatively associated with the quality of contributions. Besides this straightforward statistic, we also look at how many times user's photos have been seen by other members of the community. A user whose contributions receive increased attention from the community may be viewed as an expert photographer. We propose 3 features extracted from photo uploads and views statistics:

- *total_photos:* the number of photos a user has uploaded to Flickr.
- *avg_photo_views:* the average number of views per photo.
- *%_photos_with_at_least_100_views:* the percentage of photos that have been viewed at least 100 times. We propose this feature so that we would have an indicator for users that have a more uniform distribution of photo views. This counteracts the case in which there is a strongly skewed distribution of views which would lead to a high average from only few contributions.

5.4.2.2 Groups

In Flickr, users have the option to create groups that allow people who have similar interests to get together and share their photos. Flickr groups may form around users

sharing a common interest (brands of cars, animals etc.), or they may gather images taken with a specific camera brand or setting (black and white, light setting). This is more detailed in Chapter 3. It is also possible for a group to be created so that users coming from the same geographical location share their contributions. In a pioneering work, Negoescu et al. [335] looked at the involvement of users in groups and found, among others, that user group loyalty is generally low and most users share the same photos in different groups. We are more interested on what we could infer about a user from the groups he or she is part of. For instance, a user who is member of many groups may be more motivated to share high quality content than a user who prefers to keep his or her photos only in his or her collection. We look at the number of groups a user belongs to but also at the nature of those groups. To summarize, we extract the following features from group data:

- *groups_count:* number of groups a user is part of.
- *avg_groups_members:* the average number of members of the groups the user belongs to. A low membership may indicate more specialized groups or groups that limit the number of members. We assume that a user that belongs to many such groups may suggest a higher level of expertise.
- *avg_groups_photos:* the average numbers of photos found in the groups the user belongs to. As the number of users, we consider this feature to be a possible indicator for a group's level of specificity.

5.4.2.3 Photosets

Flickr users can organize their images in photosets, either at upload time or later, by selected a list of images to be grouped. When kept private, photosets serve the purpose to improve the organization of a user's personal collection. Public photosets give hints on a user's interest to group his or her photos for the benefit of other members of the community. A user can receive feedback for his or her photosets through comments given by other users. We propose the following features from photosets data:

- *total_photosets:* the total number of photosets created by a user.
- *photosets_avg_views:* the average number of times a photoset was viewed by other members of the community.
- *photosets_avg_comments:* the average number of comments made for a photoset by other members of the community.

5.4.2.4 Given Photo Favorites

A user can show his or her appreciation for photos of other members by marking them as favorites. We see this as another indicator for the user's involvement in the Flickr community. Here, we analyze only the number of photos a user has favorited and metadata associated to those photos. Although the number a favorites a user receives for his or her contributions may serve as a feature for credibility, Flickr does not include this information in the photos metadata and a separate API call is required for each photo individually. This renders it impractical for the immediate scope of this work. We propose the following features:

Metadata features. We include here photo, user and title words counts.

- *total_favorited_photos:* the number of photos a user marked as his favorites.
- *%_unique_users_favorited:* the number of unique users for whom the target user has favored at least one photo, divided by the total number of given favorites. Through this feature, we want to differentiate users that have a narrow circle of ties in the community for which they give favorites from those that give favorites to a more diverse set of users.
- *%_unique_words_int_favorited_titles:* the percentage of unique words found in the titles of the photos the user marked as favorites. This feature can be seen as a signal for the diversity of topics a user is interested in.

Temporal features. Sharing the same motivation as that behind proposing temporal features for a user's uploads, we consider the distribution over time of a user's favorites as a clue for his or her engagement in the community.

- *different_favorited_days:* the number of different days in which a user has marked as favorite at least one photo.
- *average_favorited_time_delay_days:* the average time elapsed between two consecutive given favorites, measured in days.
- *average_favorited_time_delay_minutes:* the average time elapsed between two consecutive given favorites, measured in minutes.

5.4.2.5 Contacts

Starting from the sample of the Flickr network we introduced in Section 5.4.2, and retain number of contacts a user has and the number of other members who have the user among

their contacts. We also investigate the use of well established link analysis algorithms, such as PageRank and HITS for estimating the credibility of Flickr users.

In the original PageRank algorithm [255] a single PageRank vector is computed using the link structure of the Web, to capture the relative importance of Web pages for the purpose of improving the ranking of search query results. PageRank has been used to analyze not only web pages but also users in networks where there is a unidirectional relationships between users (such as the *follower* relationship in Twitter) [366]. Considering that the link between two Flickr contacts is also unidirectional, we extract the PageRank score of the users in our evaluation dataset.

The HITS algorithm was developed by Kleinberg [160] and starts from the premise that web pages serve two purposes: to provide information on a topic, and to provide links to other pages giving information on a topic. HITS mines the link structure of the Web and discovers the thematically related Web communities that consist of *authorities* and *hubs*. As described in [367], authorities are the central Web pages in the context of particular query topics. For a wide range of topics, the strongest authorities consciously do not link to one another. Thus, they can only be connected by an intermediate layer of relatively anonymous hub pages, which link in a correlated way to a thematically related set of authorities. Similar to PageRank, HITS has been used beyond the scope of Web pages and can serve as a method of identifying experts in online question answering communities [148] or opinion leaders in Twitter [368]. Here, we use HITS in a similar fashion, with the goal of finding both influential users and hubs in their Flickr contact network. For a user in our dataset, we extract his or her HITS metrics but also statistics on the HITS scores of his or her immediate contacts.

In Figure 5.5, we give an example of a user's subgraph. We retain a first set of users, including user *12285897@N00* and all of his or her contacts, from our credibility dataset. Then, we select the users that have at least one user from that set among their contacts. For visualization purposes, we keep only a random subsample of the nodes, making sure to include the original user. Nodes are colored in respect to their HITS authority score. The darkest the color, the higher the authority score of that user. Similarly, the size of the labels representing users' Flickr ids are proportional to the authority score. Having the most outgoing links, the user *12285897@N00* has the highest hub score in this subgraph. We noticed that this observation does not hold for all the users in our dataset.

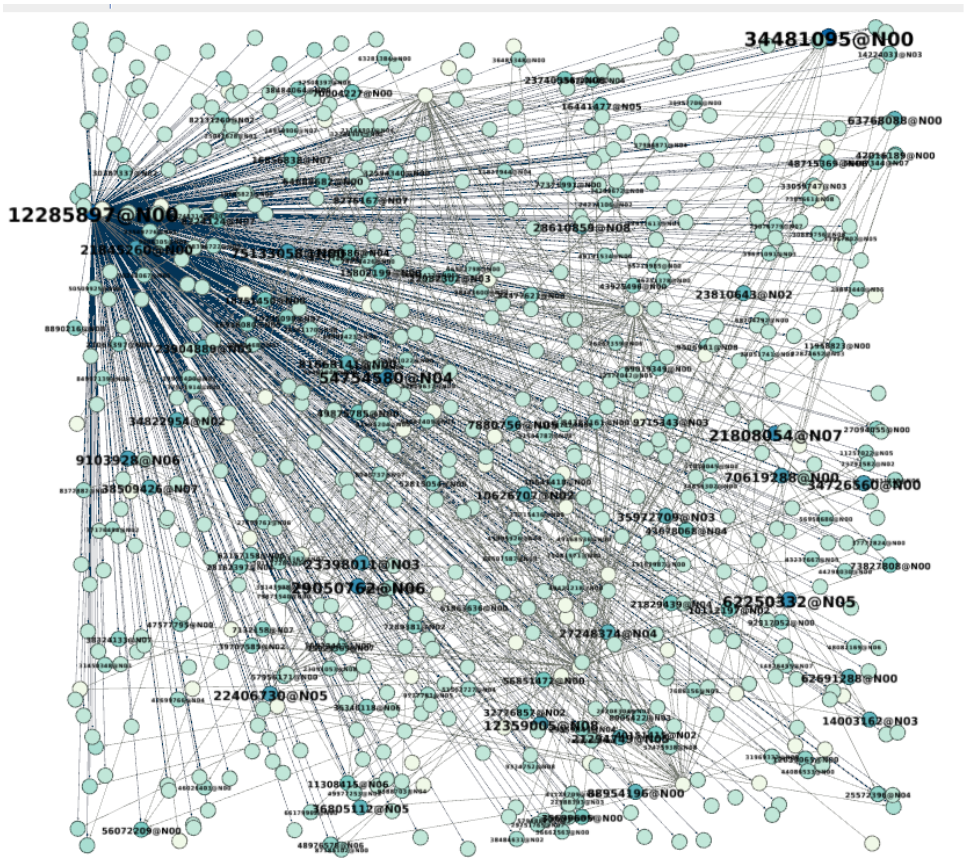


FIGURE 5.5: Example of a user's contacts subgraph. Node colors and label sizes are proportional to the HITS authority score.

By analyzing subgraphs that were extracted by starting from other users, we discovered multiple nodes having higher hub scores than the original user. When looking at the authority scores, it is even more obvious that the bias towards the users in our dataset that we introduced collecting the contacts data does not have a strong impact on the link analysis methods. In the upper right corner of Figure 5.5, we can see a user with a higher authority score than the one used to generate the network subsample. These observations hint that we can obtain reliable network metrics even if we favor a set of users when collecting contacts data. Next, we briefly present the set of features we extract from contacts network data.

User network metrics features. This is a straightforward set of features, containing the network metrics computed directly for a target user.

- *user_in_degree:* the number of users that have the target user among their contacts.
- *user_out_degree:* the number of contacts of the user.
- *user_authority:* the HITS authority score of the user.

- *user_hub:* the HITS hub score of the user.
- *user_pagerank:* the PageRank score of the user.

Contacts network metrics features. Although when computing link analysis metrics for a user, his or her contacts are implicitly taken into consideration, we propose a subset of features that directly target a user's contacts. We chose this approach so that we can have a more detailed analysis of network features for credibility estimation, considering that we fully download data only for the set of immediate contacts of a user.

- *avg_contacts_in_degree:* the average *in_degree* of the user's contacts.
- *avg_contacts_out_degree:* the average *out_degree* of the user's contacts.
- *avg_contacts_authority:* the average HITS authority score of the user's contacts.
- *avg_contacts_hub:* the average HITS hub score of the user's contacts.
- *avg_contacts_pagerank:* the average PageRank score of the user's contacts.

In summary, we extract a total of **36** context credibility estimators, covering different aspects of a user's contributions to Flickr.

5.4.3 Feature Analysis

We evaluate the features described in the previous sections by looking at how well they correlate with the manual credibility scores introduced in Section 5.3. We use Spearman's rank correlation for this purpose. The choice of Spearman correlation over Pearson is justified by our final goal of comparing an user ranking given the manual credibility score to one dictated by a user feature and not necessarily to test if there is a linear relationship between the credibility scores and the features.

From Table 5.2 we can draw the general conclusion that, taken individually, all of the proposed features are poorly correlated with the manual credibility scores. However, when comparing features, we observe that some of the hypotheses listed in the previous section are confirmed. Surprisingly, the strongest indicators for credibility are two of the photosets features (*photosets_avg_comments* and *photosets_avg_views*). Both features reveal the attention a user's contributions receive from other members of the community indicating that there is a weak positive correlation between the popularity of a user's photosets and quality of a users contributions. Note that, while the number of views of a user's photosets is the second best correlated feature, the number of photo views has

close to zero correlation with credibility. This may be explained by the fact that it is unlikely for a user with a large number of photos to have many views for the majority of them. On the contrary, a user who has a reduced number of sets can accumulate a lot of views on them. Photosets are also made with more consideration from the user and reflect his or her intention to provide curated content.

TABLE 5.2: Spearman correlation between the proposed features and the ground truth credibility scores.

Feature name	**Spearman**	**Feature name**	**Spearman**
photosets_avg_comments	0.266	avg_date_taken_upload_delay_hours	0.063
photosets_avg_views	0.202	avg_contacts_out_degree	0.053
different_upload_days	0.166	avg_photo_views	0.053
different_favorited_days	0.161	user_hub	0.049
avg_upload_time_delay_minutes	0.149	%_photos_with_at_least_100_views	0.033
total_photosets	0.116	title_vocabulary_size	0.017
avg_contacts_hub	0.114	user_out_degree	0.005
avg_contacts_in_degree	0.105	user_pagerank	0.005
avg_contacts_authority	0.105	avg_photo_taken_time_delay_days	-0.02
user_authority	0.102	%_unique_words_int_favorited_titles	-0.054
user_in_degree	0.102	avg_favorited_time_delay_minutes	-0.059
avg_photo_taken_time_delay_minutes	0.093	%_unique_users_favorited	-0.059
avg_contacts_pagerank	0.092	total_photos	-0.084
groups_count	0.092	avg_upload_time_delay_days	-0.093
total_favorited_photos	0.091	title_capitalized_words_percentage	-0.095
different_photo_taken_days	0.078	avg_favorited_time_delay_days	-0.099
avg_groups_members	0.076	avg_title_word_counts	-0.102
avg_groups_photos	0.069	title_bulk_percentage	-0.114

While the observed data precludes a definitive statement, the assumption made in Section 5.4.2 about the bulk percentage among photo titles being a good indicator for low credibility is partially supported. Although the correlation between *title_bulk_percentage* and the manual credibility scores has a low absolute value, it still presents the highest inverse correlation among all of the proposed features. In the same register as the results reported by Ye and Nov [358], who find a negative correlation between the quantity and quality of a user's contributions, we observe a negative correlation between the *total_photos* feature and the manual credibility score. Although negative, the correlation score is very small, falling close to indicating no correlation. Surprisingly, none of the proposed contact-related features seem to be strongly related to credibility. With the exception of the number of contacts, the features are extracted from a sample of the Flickr network. In spite of the fact that we tried to minimize the impact of this shortcoming, without access to the full Flickr contacts network, we cannot give a final conclusion on the usefulness of contacts features. Nevertheless, most of these features are close together in the upper half of the ranked feature list presented in Table 5.2.

Temporal features confirm our assumption about the link between the time spent by a users adding contributions in Flickr, either uploading images or giving favorites to other user's photos, and the quality of his or her contributions. We observe a positive correlation for the features referring to the number of different days a user has been active in Flickr (*different_upload_days* and *different_favorited_days*) and a negative correlation for features that relate to the length of pauses between contributions (e.g. *avg_upload_time_delay_days*, *avg_favorited_time_delay_days*).

5.5 Using visual concepts to derive a user credibility estimator

Throughout this Chapter, we examine 66 indicators for user credibility but we pay special attention to the visual cues. This aspect is proper to images and has not been previously addressed in credibility studies. In the context of image sharing platforms, we defined credibility primarily through the concept of quality. When looking at his or her contributions, we consider a Flickr user to be credible if he or she is an expert tagger who provides reliable high quality annotations for the photos shared on the platform. The quality of a tag list is objectively evaluated in regards to the content of the associated image and not the context in which the user provided the tags. In this sense, a high quality tag set is one in which the individual tags can be identified in the image and can be easily be deemed relevant by other users, thus being useful not only for the uploader but also for the entire community and can be correctly indexed by a tag based image retrieval system.

5.5.1 Visual credibility estimator extraction

In this section, we propose a credibility feature that directly exploits the image-tag associations of a user. In order to achieve this goal, we use the large collection on visual concept models introduced in Chapter 3. The intuition behind this approach is that we place our confidence in the learned models to provide an accurate global view on the visual representation of a certain concept. We use two sets of visual classifiers (trained on ImageNet concepts and Flickr groups) that were learned from Overfeat descriptors.

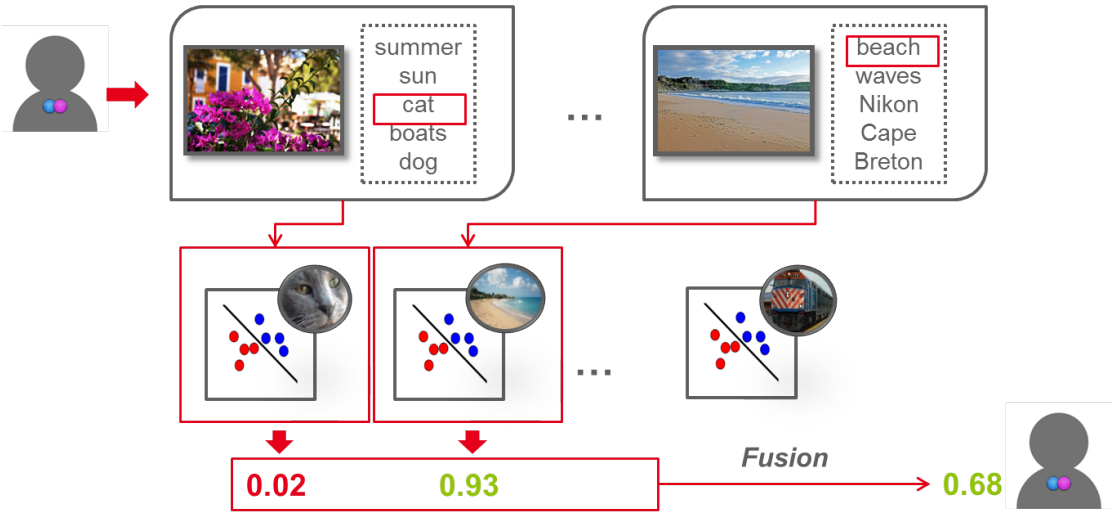


FIGURE 5.6: Visual credibility estimator extraction framework.

In Figure 5.6, we give an example of the extraction process for the proposed visual credibility feature. When being confronted with a tag given by the user, we classify the image with the model for the that tag and consider the obtained probability score as the indicator of the quality of the tag-image association. To obtain an average user score, we tested three *fusion* methods: the mean, maximum or minimum prediction scores. The results of this evaluation are presented in Tables 5.3 and 5.4.

Next, we compare visual credibility estimates extracted from predictions of ImageNet visual models with those of visual model trained from Flickr Groups. For each user in *Div150Cred*, we downloaded at most 300 images whose textual annotations match at least one ImageNet concept and 300 images whose textual annotations match the most representative tag of a Flickr group (for the textual representation of Flickr groups, see Chapter 3. In the case of *MTTCred*, we downloaded at most 1000 images per user. Flickr annotations are selected either from tags or from the image title and are all referred as tags hereafter.

A tag can correspond to more than one ImageNet concept, as shown in Figure 5.7. Similarly, a tag can be associated with several Flickr groups. In the case of ImageNet based concepts, we extract the prediction score for each match. For Flickr groups, we can have over 100 visual models corresponding to a tag. In these cases, we extract predictions only for the top 10 Flickr group models ranked by their cross-validation scores. A tag cannot visually represent different concepts in an image. Therefore, for each tag that corresponds to more than one visual model (either ImageNet or Flickr group), we keep only the maximum prediction score for that tag.

Both in Table 5.3 and Table 5.4, we can see that the differences of correlation coefficients are small between visual credibility features obtained from ImageNet and Flick

TABLE 5.3: Spearman correlation between visual credibility estimators obtained with different individual tag prediction scores *fusion* strategies and the ground truth user credibility scores on the *MTTCred* dataset. Results are shown both for ImageNet and Flickr groups visual models.

	Visual concept collection	
Visual credibility extraction	*ImageNet*	*Flickr groups*
visual_cred_min	0.208	0.131
visual_cred_max	0.314	0.328
visual_cred_mean	0.346	**0.362**

TABLE 5.4: Spearman correlation between visual credibility estimators obtained with different individual tag prediction scores *fusion* strategies and the ground truth user credibility scores on the *Div150Cred* dataset. Results are shown both for ImageNet and Flickr groups visual models.

	Visual concept collection	
Visual credibility extraction	*ImageNet*	*Flickr groups*
visual_cred_min	0.173	0.148
visual_cred_max	0.302	0.297
visual_cred_mean	**0.356**	0.346

groups. This is a somewhat surprising result, given that Flick group visual models perform better the ImageNet ones when used for image retrieval (see Chapter 4). Given that for both evaluation datasets and for both sets of visual models, averaging the predictions for each image leads to features that are highest correlated with the ground truth credibility scores, in the following Sections we will simply note *visual_credibility* the *visual_cred_mean* estimator.

The extraction of the visual credibility estimate for user u can be summarized by the following Equation:

$$visual\_credibility(u) = \frac{\sum_{I \in C_u} \frac{\sum_{t \in T_I} W^t f^I}{|T_I|}}{|C_u|} \tag{5.1}$$

, where W^t is the vector model weights for the tag t and f^I is the Overfeat representation of image I. We have described the visual model building in Chapter 3.

5.5.2 Discussion

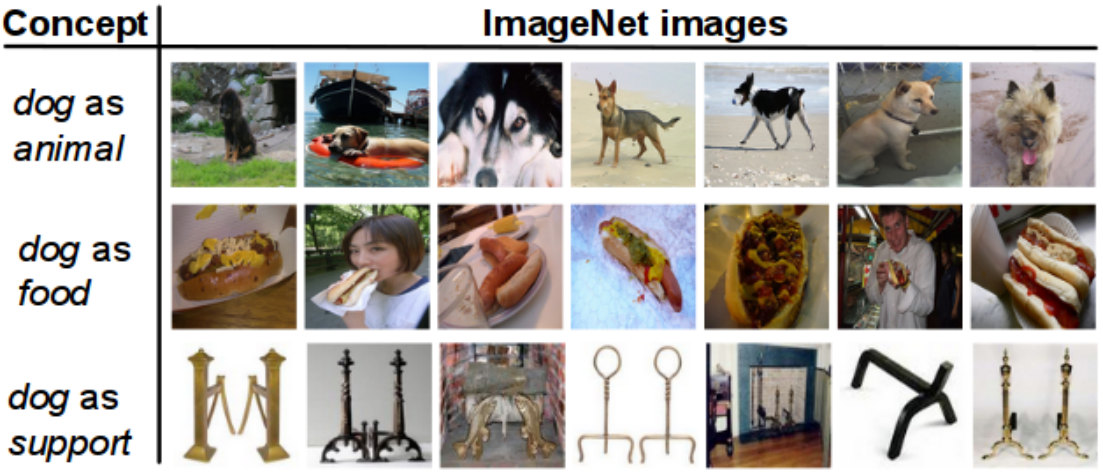


FIGURE 5.7: Different encounters of the word *dog* among ImageNet concepts.

Here, we discuss possible problems that we may encounter in our visual credibility feature extraction framework. We focus on visual models built on top of unambiguous ImageNet concepts and tested for Flickr annotations, which are often ambiguous. For instance, if an unknown image annotated with *dog* is tested, which of the three senses of *dog* modeled in ImageNet (Figure 5.7) should be used? An inspection of Flickr results shows that most images annotated with *dog* depict *animals* but there are some of them which depict *dog* as *food* and *dog* as *support*. Our credibility estimator needs to be able to automatically select the right sense of *dog* for the content of the tested image. As presented above, we opted to compare the tag-image pair to all models available for the tag and retain only the maximum classification score. Preliminary tests showed that this procedure has good behavior and it is thus used in the experiments.

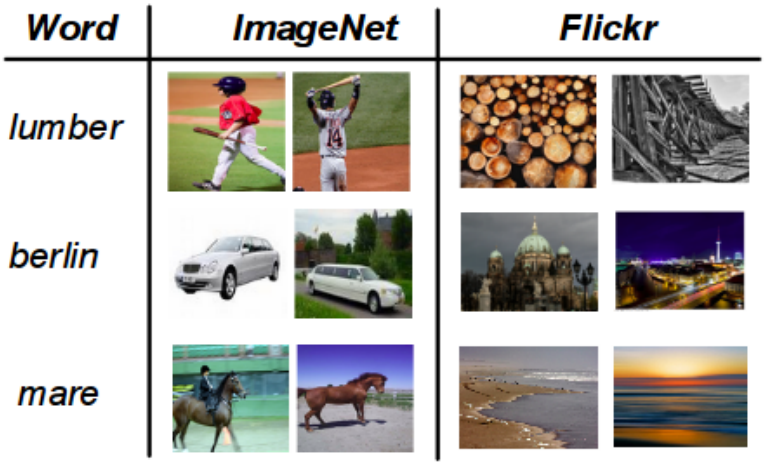


FIGURE 5.8: Word mismatch between ImageNet concepts and Flickr groups.

Beyond ambiguity, another problem is the coverage of ImageNet, with some important senses of words not being included. For instance, *berlin* is represented as *car* but not as *city*. Yet another problem is the fact that ImageNet is built for English concepts while Flickr annotations are often performed in other languages. For instance, *mare* is represented as *female horse* in ImageNet while the dominant Flickr sense is the Italian translation of *sea*. These problems should represent limitations of our method when using ImageNet models. However, the results presented in Tables 5.3 and 5.4 show that these limitations of ImageNet models do not have a high impact on the visual credibility features, when compared to features built from Flickr group models. As we will show in the following Sections, as well as in Chapter 6, the visual credibility feature introduces in this Section proves to be the best individual estimator for manual user tagging credibility scores.

5.6 Content features as credibility estimators

We provide in this section a list of features that can be extracted directly from the main content of a user's contributions on Flickr (*i.e.* images and tags). Similar to the previous Section, we first detail the feature extraction process and evaluate the proposed credibility estimators using the Spearman correlation with the ground truth credibility score. Besides testing on the *MTTCred* dataset, we also extract content features for the *DIV150Cred* dataset.

5.6.1 Data acquisition

Tags, alongside the actual image, are the main content produced by a Flickr user. For most of the tag based features detailed in this Section, we require tag frequency and co-occurrence statistics from a large sample of Flickr images. To obtain a representative set of tag lists, we first gather a collection of Flickr images metadata by download information for 50 000 Flickr groups (the Flickr group collection presented in Chapter 3). We first eliminate bulk tagging and obtain a set of 20 737 794 unique tag set. Then, we *clean* the tag lists in order to reduce the level of noise inherent to Flickr tags. We eliminate tags that contain numbers and special characters and tags that have less than 3 and more than 15 characters. We finally get a list of 3 952 087 unique tags and we sort this list by the number of tag occurrences. For instance, the most frequent three tags are: *canon* (2 214 241 instances), *nikon* (2 159 691 instances) and *nature* (1 840 668 instances). From the initial tanked list, we keep only the top 100 000 most frequent tags. We note this ranked list as T_D. This way, we reduce computation cost without loosing many relevant tags (the last 3 kept tags are *greenmount*, *sketchbookstyle*, and *sher*, each with

220 counts). We then build a matrix with co-occurrence counts for the top 10 000 most frequent tags. To decrease computing complexity for the extraction of co-occurrence based credibility estimates, we reduce the set of tags used in the co-occurrence matrix to 10% of the initial ranked tag list. In the following, we will refer this matrix as Mc.

We represent each tag through a term frequency–inverse document frequency (*tf-idf*) language model. We treat each line in Mc (*i.e.* the co-occurrence counts with the other concepts), as a document in the classical understanding of the *tf-idf* model. To get the final representation for tag t_i, we normalize each value in the corresponding line from M_c (noted $Mc[i]$) by *idf*. For each j in $[1, |Mc[i]|])$, we have:

$$tf - idf(t_j) = Mc[i][j]) \times \log(\frac{|T_D]}{f(t_j, T_D)}) \tag{5.2}$$

,where $Mc[i][j]$ is the co-occurrence value between t_i and t_j, and $f(t_j, T_D)$ the frequency of tag t_j in the list T_D.

By applying Equation 5.2 to all values in Mc we obtain the *tf-idf* normalized matrix M_{tfidf}^{global}. This matrix represents the global language model for tags. In the following Subsection, we define a user specific *tf-idf* language model.

5.6.2 Feature extraction

Visual content. We extract the visual credibility estimator examined in detail in Section 5.5.

visual_credibility: Given the initial tests performed in Section 5.5, we use the feature *visual_cred_mean* (*i.e.* averaging the visual model prediction for a user's visual credibility score for each image). Also, following the observations from Section 5.5, for the *MTTCred* dataset we use the visual credibility estimator built upon Flickr group visual models, while for *Div150Cred* we use the one built on top of ImageNet visual concept models. Throughout this chapter, for each dataset, we will simply use *visual_credibility*, for the visual credibility estimator, without reiterating the underlying extraction configuration.

Tags text quality. These features are a representation of the text quality component of credibility, as presented in Section 2.2.1.3 of Chapter 2. We extract the following features:

- *tags_with_numbers_percentage*: The percentage of tags that contain numbers. In most cases, these tags are not relevant to the visual content of the image.

- *tags_non_alpha_percentage*: The percentage of tags that contain non alpha-numeric characters.
- *tags_len_over_10_percentage*: The percentage of tags that have over 10 characters.
- *tags_len_over_15_percentage*: The percentage of tags that have over 15 characters.
- *vocabulary_size*: The number of unique tags.

Tags counts. We extract the following features: We note by t_{rank} the rank in of t in T_D and by t_{freq} the number of times t appears in the large collection from which we build T_D. The features based in these scores represent an indicator to whether a user prefers to use more specific or generic tags. We extract the following features:

- *avg_min_tag_rank*: This feature is obtained by averaging the $min\_rank_(T_I)$ values for all the tags lists u_{T_I} of a user. $min\_rank_(T_I)$ is the minimum t_{rank} value of all tags t in T_I.
- *avg_max_tag_rank*: This feature is obtained by averaging the $max\_rank_(T_I)$ values for all the tags lists u_{T_I} of a user. $max\_rank_(T_I)$ is the maximum t_{rank} value of all tags t in T_I.
- *avg_mean_tag_rank*: This feature is obtained by averaging the $mean\_rank_(T_I)$ values for all the tags lists u_{T_I} of a user. $mean\_rank_(T_I)$ is the mean t_{rank} value of all tags t in T_I.
- *avg_min_tag_freq*: This feature is obtained by averaging the $min\_freq_(T_I)$ values for all the tags lists u_{T_I} of a user. $min\_freq_(T_I)$ is the minimum t_{freq} value of all tags t in T_I.
- *avg_max_tag_freq*: This feature is obtained by averaging the $max\_freq_(T_I)$ values for all the tags lists u_{T_I} of a user. $max\_freq_(T_I)$ is the maximum t_{freq} value of all tags t in T_I.
- *avg_mean_tag_freq*: This feature is obtained by averaging the $mean\_freq_(T_I)$ values for all the tags lists u_{T_I} of a user. $mean\_freq_(T_I)$ is the mean t_{freq} value of all tags t in T_I.
- *tags_top_10k_percentage*: The percentage of a user's tags that are in the first 10_000 tags from T_D.

- *tags_top_50k_percentage*: The percentage of a user's tags that are in the first 50_000 tags from T_D.
- *avg_tags_per_photo*: The average number of tags for that a user puts for each photo.
- *tag_counts_mean*: We count the number of times each a users puts each tag for an image. This descriptor is the average use of a user's tag. It represents another way of determining if the user prefers to use the same tags for his/her photos or uses a diverse set of tags.
- *tag_counts_stdev*: The standard deviation of the list of a user's tag usage counts.
- *bulk_percentage*: The percentage of tag list that are used at least two times among the full set of tag lists for a user.

Tags language model. Through this feature family, we aim to capture the similarity between a community tag language model and a user specific tag language model. Our intention here is to exploit the 'wisdom of crowds' and hypothesis that a user has a higher tagging credibility if his/her tag models are closer to those of the community. In the previous section, we defined M_{tfidf}^{global}, the matrix of language model for 10,000 tags. For each user u, we extract a specific tag language model (M_{tfidf}^{u}) using Equation 5.2. The single difference is that we replace the global Mc co-occurrence matrix, with a user specific one (M_c^u) in this Equation. Let u_{T_I} be the lists of image tags for user u. M_c^u is obtained by counting the tag co-occurrences from u_{T_I}.

Let T_{rank} be the set of 10,000 tags found in M_{tfidf}^{global}. We test two similarity measures between a tag's user specific language model and the tag's global language model: dot product and cosine. For each tag t_i from u_{T_I} that is in T_{rank}, we have:

$$product(t_i) = M_{tfidf}^{global}[i] \bullet M_{tfidf}^{u}[i] \tag{5.3}$$

$$cosine(t_i) = cos(M_{tfidf}^{global}[i], M_{tfidf}^{u}[i]) = \frac{M_{tfidf}^{global}[i] \bullet M_{tfidf}^{u}[i]}{\parallel M_{tfidf}^{global}[i] \parallel \parallel M_{tfidf}^{u}[i] \parallel} \tag{5.4}$$

We extract the following tag language model based features:

- *avg_mean_product*: This feature is obtained by averaging the $mean\_product(T_I)$ values for all the tags lists u_{T_I} of a user. $mean\_product(T_I)$ is the mean $product(t_i)$ value of all tags t_i in $T_I \cap T_{rank}$.

- *avg_max_product*: This feature is obtained by averaging the $max\_product(T_I)$ values for all the tags lists u_{T_I} of a user. $max\_product(T_I)$ is the maximum $product(t_i)$ value of all tags t_i in $T_I \cap T_{rank}$.

- *avg_min_product*: This feature is obtained by averaging the $min\_product(T_I)$ values for all the tags lists u_{T_I} of a user. $min\_product(T_I)$ is the minimum $product(t_i)$ value of all tags t_i in $T_I \cap T_{rank}$.

- *avg_mean_cosine*: This feature is obtained by averaging the $mean\_cosine(T_I)$ values for all the tags lists u_{T_I} of a user. $mean\_cosine(T_I)$ is the mean $cosine(t_i)$ value of all tags t_i in $T_I \cap T_{rank}$.

- *avg_max_cosine*: This feature is obtained by averaging the $max\_cosine(T_I)$ values for all the tags lists u_{T_I} of a user. $max\_cosine(T_I)$ is the maximum $cosine(t_i)$ value of all tags t_i in $T_I \cap T_{rank}$.

- *avg_min_cosine*: This feature is obtained by averaging the $min\_cosine(T_I)$ values for all the tags lists u_{T_I} of a user. $min\_cosine(T_I)$ is the minimum $cosine(t_i)$ value of all tags t_i in $T_I \cap T_{rank}$.

Image tag clarity (ITC). Modified version of the *ITC* score introduced in [369], who use visual words w do build *language models*. Instead, we use a classical textual approach of a language model (*term frequency (tf)*). In our *ITC* implementation, w is a tag associated to an image. Let C_t be the set of images annotated by a tag t. The image tag clarity score of t, denoted by $ITC(t)$, is defined by Sun and Bhowmick [369] as the *KL-divergence* between the *tag language model* ($P(w|C_t)$) and the *collection language model* ($P(w|D)$). It is expressed by the following equation:

$$ITC(t) = KL(C_t||D) = \sum_w P(w|C_t) log_2 \frac{P(w|C_t)}{P(w|D)} \tag{5.5}$$

The collection language model is estimated by the word frequency in the collection. Sun and Bhowmick [369] propose two methods to estimate the tag language model: either have a unified representation for $P(w|C_t)$ or weigh each image in C_t with a *centrality function*. We chose the first method and treat all images as equally representative of a tag t. Therefore according to Sun and Bhowmick [369], we have:

$$P(w|C_t) = \sum_{I \in C_t} \frac{1}{|C_t|} P_{lm}(w|I) \tag{5.6}$$

In our case:

$$P_{lm}(w|I) = \begin{cases} 1, \text{if I is tagged with } w \\ 0, \text{otherwise} \end{cases} \tag{5.7}$$

Simply put, $P(w|C_t)$ is the frequency of tag w in the C_t collection.

We extract the following ITC based features:

- avg_mean_itc: This feature is obtained by averaging the $mean\_itc(T_I)$ values for all the tags lists u_{T_I} of a user. $mean\_itc(T_I)$ is the mean $ITC(t)$ value of all tags t in T_I.
- avg_max_itc: This feature is obtained by averaging the $max\_itc(T_I)$ values for all the tags lists u_{T_I} of a user. $max\_itc(T_I)$ is the maximum $ITC(t)$ value of all tags t in T_I.
- avg_min_itc: This feature is obtained by averaging the $min\_itc(T_I)$ values for all the tags lists u_{T_I} of a user. $min\_itc(T_I)$ is the minimum $ITC(t)$ value of all tags t in T_I.

Pointwise mutual information (PMI). We first compute the pointwise mutual information (pmi) for any pair of tags from a tag list, according to Equation 5.8.

$$pmi(t_i, t_j) = \log \frac{p(t_i, t_j)}{p(t_i)p(t_j)} \tag{5.8}$$

, where T_I is a list of tags associated with an image I, $p(t_i)$ is the probability that the tag t_i appears in our tag list collection and $p(t_i, t_j)$ is the probability that t_i and t_j appear together.

We extract the following pmi based features:

- avg_mean_pmi: The final feature is obtained by averaging the $mean\_pmi(T_I)$ (Equation 5.9) values for all tag lists u_{T_I} of a user. This feature serves as an indicator to whether a user's tagging behavior is similar to or diverges from that

of a large sample of the Flickr community.

$$mean\_pmi(T_I) = \frac{\sum_{t_i \in T_I} \sum_{t_j \in T_I \setminus t_i} \log \frac{p(t_i,t_j)}{p(t_i)p(t_j)}}{|T_I|} \tag{5.9}$$

- *avg_max_pmi*: This feature is obtained by averaging the $max\_pmi(T_I)$ values for all the tag lists u_{T_I} of a user. $max\_pmi(T_I)$ is the maximum $pmi(t_i, t_j)$ value for any (t_i, t_j) pairs of T_I.
- *avg_min_pmi*: This feature is obtained by averaging the $min\_pmi(T_I)$ values for all the tag lists u_{T_I} of a user. $min\_pmi(T_I)$ is the minimum $pmi(t_i, t_j)$ value for any (t_i, t_j) pairs of T_I.

In summary, we extract a total of **30** content credibility estimators. Although some features may seem redundant (*i.e.* extracting the mean, maximum and minimum values for a type of feature, we will show in the following Section that automatic feature selection methods benefit from a higher number of features.

5.6.3 Feature Analysis

We evaluate the features described in the previous sections by looking at how well they correlate with the manual credibility scores introduced in Section 5.3 for the *MTTCred* dataset. We also evaluate on the *Div150Cred* dataset that is presented in the same Section. We use Spearman's rank correlation for this purpose.

From Tables 5.5 we can draw the general conclusion that the content features are better correlated with the manual credibility scores than the context features introduced in the previous Section.

When comparing features, we observe that some of the hypotheses listed in the previous section are confirmed. For instance, the users users who tend to use more common tags are less likely to be deemed credible. The *avg_max_tag_rank* feature is the most negatively correlated with the ground truth credibility scores on the *Div150Cred* dataset. Also, as expected, bulk tagging is an indicator for low credibility.

TABLE 5.5: Spearman correlation between the proposed content features and the ground truth credibility scores on the *MTTCred* dataset.

Feature name	**Spearman**	**Feature name**	**Spearman**
visual_credibility	0.362	avg_min_tag_rank	0.054
avg_mean_product	0.236	avg_min_itc	0.044
avg_min_product	0.226	tags_non_alpha_percentage	0.024
avg_max_product	0.192	avg_max_tag_freq	-0.005
avg_min_cosine	0.171	tags_top_10k_percentage	-0.041
avg_min_pmi	0.169	avg_mean_tag_rank	-0.041
avg_min_tag_freq	0.164	tags_top_50k_percentage	-0.070
avg_mean_tag_freq	0.126	avg_max_pmi	-0.099
tags_len_over_10_percentage	0.125	avg_mean_itc	-0.151
tags_len_over_15_percentage	0.123	avg_tags_per_photo	-0.168
avg_mean_cosine	0.120	avg_max_tag_rank	-0.171
tags_with_numbers_percentage	0.100	tag_counts_stdev	-0.175
avg_mean_pmi	0.082	avg_max_itc	-0.189
avg_max_cosine	0.069	bulk_percentage	-0.221
vocabulary_size	0.064	tag_counts_mean	-0.235

TABLE 5.6: Spearman correlation between the proposed content features and the ground truth credibility scores on the *Div150Cred* dataset.

Feature name	**Spearman**	**Feature name**	**Spearman**
visual_credibility	0.356	avg_min_tag_rank	0.007
avg_min_tag_freq	0.236	tags_with_numbers_percentage	0.003
avg_mean_product	0.185	bulk_percentage	-0.030
avg_max_product	0.179	tags_len_over_10_percentage	-0.035
avg_min_product	0.161	avg_max_tag_freq	-0.042
tags_top_10k_percentage	0.161	tags_len_over_15_percentage	-0.063
avg_mean_cosine	0.146	tag_counts_mean	-0.082
avg_min_cosine	0.141	avg_max_pmi	-0.085
avg_min_pmi	0.137	tag_counts_stdev	-0.129
tags_top_50k_percentage	0.133	vocabulary_size	-0.164
avg_max_cosine	0.125	avg_mean_itc	-0.169
avg_mean_tag_freq	0.121	avg_mean_tag_rank	-0.191
avg_mean_pmi	0.088	avg_tags_per_photo	-0.223
avg_min_itc	0.026	avg_max_itc	-0.235
tags_non_alpha_percentage	0.017	avg_max_tag_rank	-0.289

When comparing the correlation scores from Table 5.5 and Table 5.6, we first observe

the consistence of the *visual_ credibility* estimator. In both cases, it is the highest correlated feature and also has similar correlation scores (0.362 for *MTTCred* and 0.356 for *Div150Cred*). For the remaining features, although the ranking is does not exactly match between the two datasets, there are similarities on the usefulness of different feature types. For instance, for both datasets, the tag language model based estimators can be found among the top five ranked features. From Table 5.5, we can observe that among the highest correlated features are those that compare a user's own tag language model with global language models (*product, cosine, pmi*). This confirms our hypotheses that we can exploit *the wisdom of crowds* for deriving credibility estimates. A user is more likely to be credible if his or her tagging practices are closer to those of the community. This observation also stands for *Div150Cred* (Table 5.6).

5.7 Feature evaluation and ranking

In this Section, we investigate which features or groups of features are more informative for user credibility and how we can use regression models to learn better user credibility estimates. Our goal is to extract user credibility estimates that can be used to filter or rerank a list of retrieved items according to the users who produced them. We are interested in a fine-grained score that will allow us to rank users based on their credibility estimates. Considering that we have a single list of ground truth credibility scores, learning to rank approaches are not feasible. Given this limitation, we are not able to directly predict a ranking of users. In order to by-pass this limitation, we treat the credibility score as a continuous variable (Y) and model it a regression problem in which we fit a model that learns to approximate the credibility score:

$$Y \approx f(X, \beta)$$

, where X is the feature vector and β are the model weights. We then used the regression model to predict credibility scores and rank users according to the predictions.

Unlike classical regression problems, our final goal is not to provide an approximation of the credibility score but to rank users according to their credibility estimates. This makes evaluation metrics usually used in regression problems (e.g. the mean squared root error) uninformative for our specific task. We directly evaluate the ranking obtained from the predicted scores with that given by the manual credibility scores. Following a similar procedure used to evaluate individual features, we use the Spearman rank correlation measure to test a new ranking. When comparing multiple classifiers, we are interested in the one that maximizes the correlation between the manual rank and the predicted

rank:

$$\arg\max_m Spearman(Y^m_{pred}, Y_{man})$$

Y^m_{pred} is the prediction vector corresponding to model m. The same evaluation measure is used when comparing different feature subsets selected to train the same classifier.

5.7.1 Feature family prediction performance

We observed low correlation scores on the *MTTCred* dataset for most of the context features presented in Section 5.4. For the *Div150Cred* dataset we extracted only the metadata set of features among context features, alongside the content features. In summary, in this Section, we experiment with 66 features (30 content features and 36 context features) on the *MTTCred* dataset and with 45 features (30 content features and 15 context features) on the *Div150Cred* dataset.

When building the training set for the regression experiments, we encounter a few cases of missing values. These may be caused by technical problems or bad responses from the Flickr API, by an user who removed or made private a part of his or her data. We first address this issue by imputing the missing values using the mean value of the respective feature. Missing values account for less than 1% of our data. Due to large differences of magnitude between features, we then perform a L2 normalization. Although this does not affect ensemble models, it has a strong impact on the ability to learn of linear models. For predicting the credibility score, we test 9 models coming from 3 families of approaches:

- *linear models:* Linear Regression (LR), Ridge (linear least squares with L2 regularization), Lasso (linear Model trained with L1 prior as regularizer), Elastic Net (EN) (linear regression with combined L1 and L2 priors as regularizer), Lars (Least Angle Regression model).
- *support vector machines:* SVR (epsilon-Support Vector Regression with rbf kernel)
- *ensemble models:* Extra Trees Regressor (ETR), Random Forest Regressor (RFR), Gradient Boosting Regressor (GBR).

Due to the small size of our dataset, we evaluate each model in a leave-one-out-cross-validation (LOOCV) fashion. We select each time a different user and train a model on the remaining users. We do this for all the users in our evaluation dataset, keeping the prediction for the test user. Finally, we compare the predictions vector to the manual credibility scores. For each model, we tune the parameters on a randomly selected

validation set in which we put 10% of the users in our dataset. We consider as baseline the best individual feature in terms of Spearman correlation (in absolute value).

TABLE 5.7: Model and context features comparison for predicting a user credibility score. Results are reported in terms of the Spearman correlation between the predicted scores and the manual credibility scores on the *MTTCred* dataset. We consider as baseline the best individual feature from each feature family in terms of Spearman correlation (in absolute value).

Model	**Feature Family**					
	Metadata	*Contacts*	*Favorites*	*Groups*	*Photosets*	*All context*
LR	0.144	0.107	0.116	-0.064	0.192	0.149
Ridge	0.145	0.108	0.116	-0.064	0.192	0.15
Lasso	0.151	0.105	0.126	-0.123	0.177	0.188
EN	0.145	0.108	0.116	-0.064	0.192	0.15
Lars	0.146	0.075	0.11	-0.064	0.192	0.138
SVR	0.161	0.102	0.109	031	0.179	0.178
ETR	0.152	0.03	0.092	0.052	0.166	0.321
RFR	0.164	0.053	0.071	0.031	0.23	0.326
GBR	**0.184**	0.056	0.1	0.039	**0.283**	**0.345**
Baseline	0.166	**0.114**	**0.161**	**0.092**	0.266	0.266

TABLE 5.8: Model and content features comparison for predicting a user credibility score. Results are reported in terms of the Spearman correlation between the predicted scores and the manual credibility scores on the *MTTCred* dataset. We consider as baseline the best individual feature from each feature family in terms of Spearman correlation (in absolute value).

Model	**Feature Family**		
	Content (textual)	*Content (textual + visual)*	*Content + Context*
LR	**0.366**	**0.404**	0.382
Ridge	0.365	**0.404**	0.382
Lasso	0.352	0.394	0.380
EN	0.365	**0.404**	0.384
Lars	0.167	0.239	0.375
SVR	0.342	0.385	0.379
ETR	0.289	0.380	0.438
RFR	0.306	0.346	0.476
GBR	0.292	0.363	**0.483**
Baseline	0.236	0.362	0.362

TABLE 5.9: Model and content + context features comparison for predicting a user credibility score. Results are reported in terms of the Spearman correlation between the predicted scores and the manual credibility scores on the *Div150Cred* dataset. We consider as baseline the best individual feature from each feature family in terms of Spearman correlation (in absolute value).

Model	**Feature Family**			
	Metadata	*Content (textual)*	*Content (textual + visual)*	*Content + Metadata*
LR	**0.248**	0.362	0.409	0.398
Ridge	**0.248**	0.362	0.409	0.398
Lasso	**0.248**	0.366	**0.417**	0.403
EN	**0.248**	0.362	0.409	0.399
Lars	0.235	0.186	0.235	0.243
SVR	**0.248**	0.359	0.403	0.394
ETR	0.212	**0.371**	0.397	0.404
RFR	0.225	0.348	0.389	0.402
GBR	0.225	0.344	0.390	**0.415**
Baseline	0.179	0.289	0.356	0.356

Looking at the results presented in Tables 5.7, 5.8 and 5.9, we can observe that ensemble models outperform all other models regardless when use all of the proposed features for training. In fact, for context features on *MTTCred* (Table 5.7) they are the only one that manage to rise above the baseline. In a recent paper comparing 179 classifiers from 17 families over 121 datasets [328], Fernandez-Delgado et al. find that the family of features that gives the best performances on average over all the datasets is the ensemble family and the best individual classifier is Random Forests. We notice a similar behavior, with the exception that in our case, Random Forest is not the best performing model. The best configuration is given when a Gradient Boosting Regressor (GBR) model is trained on the full feature set. In this case, we observe a 30% relative improvement over the best individual feature (*i.e. photosets_avg_comments*).

In Table, 5.7, when comparing individual feature families, we observe that only when training a model on metadata and photosets feature families, we obtain a correlation score higher than the one given by the best individual features from each family. This result is not surprising, considering that the best features come from the photosets feature family. In total, we get only five configurations in which the baseline correlation score for an individual feature family is surpassed. In Tables 5.8 and 5.9, when comparing models trained only on content features, surprisingly, with a single exception (ETR trained with textual content features on *Div150Cred*), the linear models outperform the

ensemble ones. This may be explained by the fact that content features are less effected by normalization than context features (which have a more expanded range of values). Thus, we may get a loss of discrimination capability when normalizing context features. Given that ensemble models have identical performance with and without normalization, this can explain why ensemble models outperform the linear ones when adding context features.

Although in the previous section we saw that individual features are poorly correlated with the manual credibility scores, we are able to learn a regression model that clearly offers a better credibility estimate than any of the features. This result validates the use of regression models for predicting a score which is later used for ranking. Also, we can see that content features are overall better than context features. On both datasets, we can observe an improvement on the tag based content features when adding just the visual credibility feature. Surprisingly, on *Div150Cred*, adding metadata features to content features lowers the correlation of the predictions for most linear models.

5.7.2 Feature importance

In Section 5.4.3, we looked at the correlation between individual features and the manual credibility scores. We propose here two other methods for analyzing the usefulness of features in estimating credibility scores. The first one is given by a property of tree ensemble methods in which the depth of a feature used as a decision node in a tree can be used to assess the relative importance of that feature with respect to the predictability of the target variable. We extract the feature importance from the learned GBR model. The second one represents a feature ranking given by the weights learned by a linear model. For this we chose the Linear Regression (LR) model.

In Figure 5.9, we provide the ranked list of 66 features according to their role in training the GBR model on *MTTCred*, while on Figure 5.10, we present the features ranked according to the feature importance score provided by the LR model on *MTTCred*. Similarly, in Figure 5.11, we give the ranked list of 45 features according to their role in training the GBR model on *Div150Cred*, while in Figure 5.12, we present the features ranked according to the feature importance score provided by the LR model on *Div150Cred*. In all four plots, we normalize the importance scores by giving the most important feature a score of 100 and then relating other feature to it.

In Figure 5.9, we observe that only 9 features out of 35 have an importance score higher than 50% of the score associated with the best feature. This indicates that selecting only top features for training may improve the predicted scores. We compare this ranking to the one introduced by the Spearman correlation score, shown in

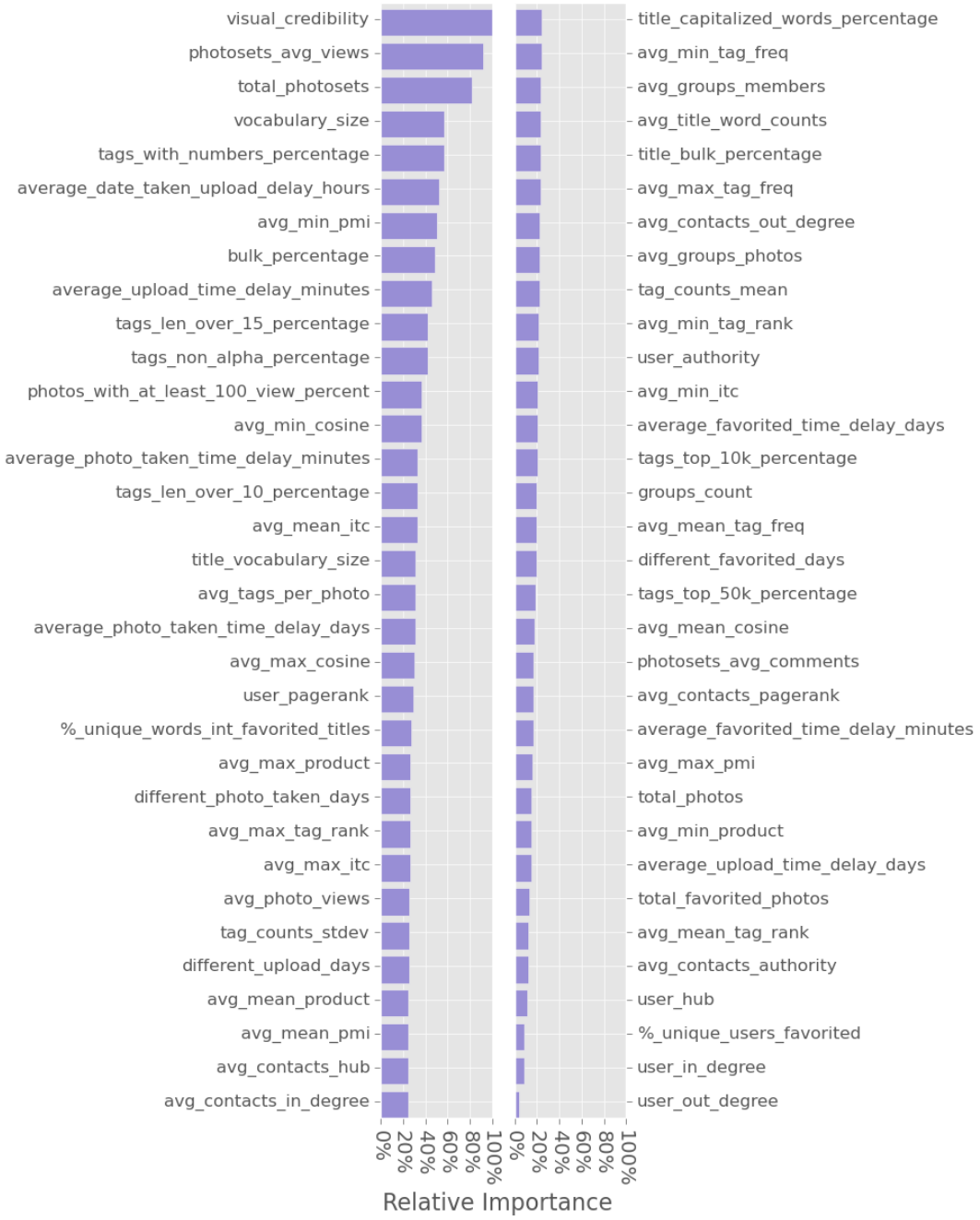


FIGURE 5.9: Features ranked according to the feature importance score provided by the GBR model on *MTTCred*.

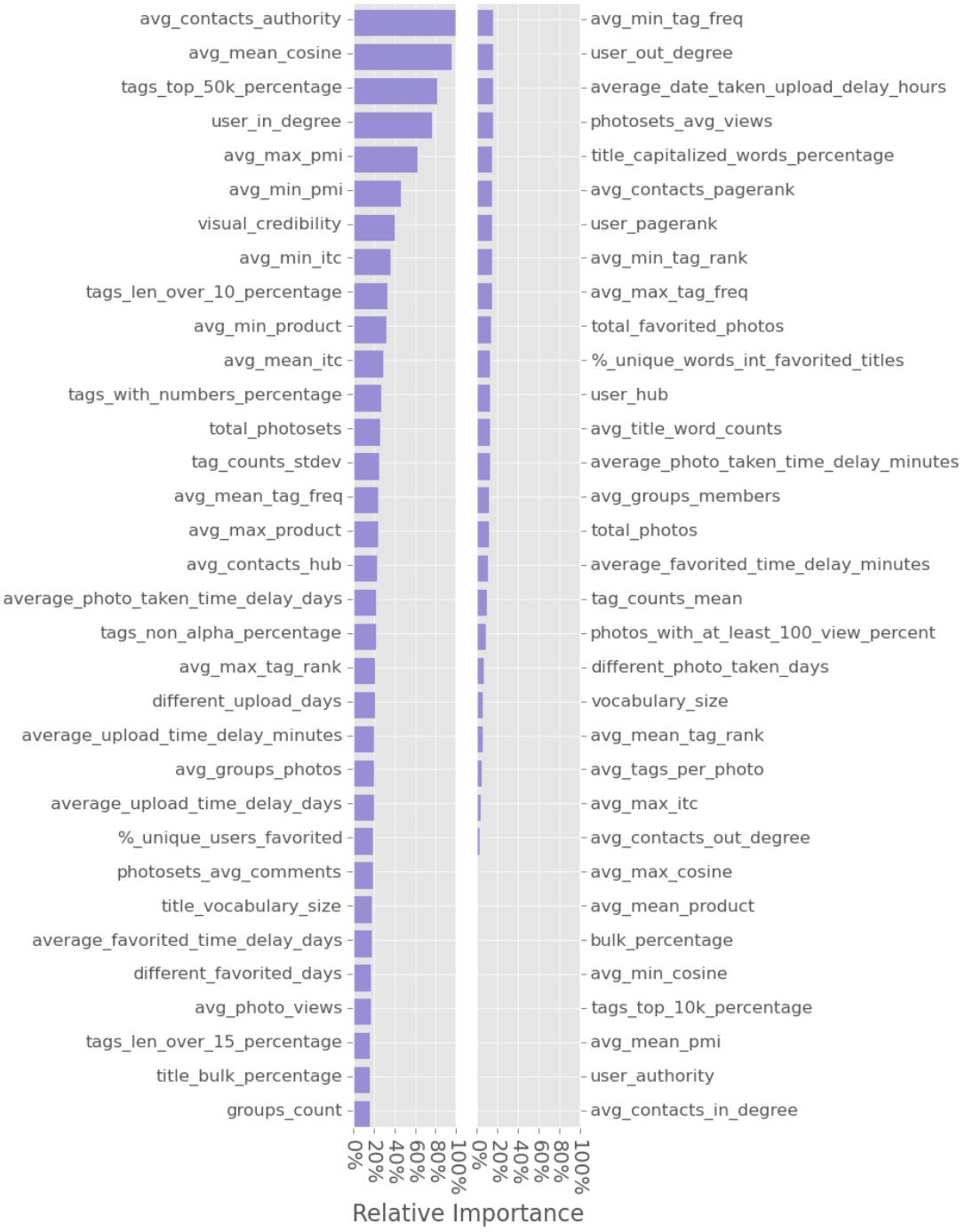


FIGURE 5.10: Features ranked according to the feature importance score provided by the LR model on *MTTCred*.

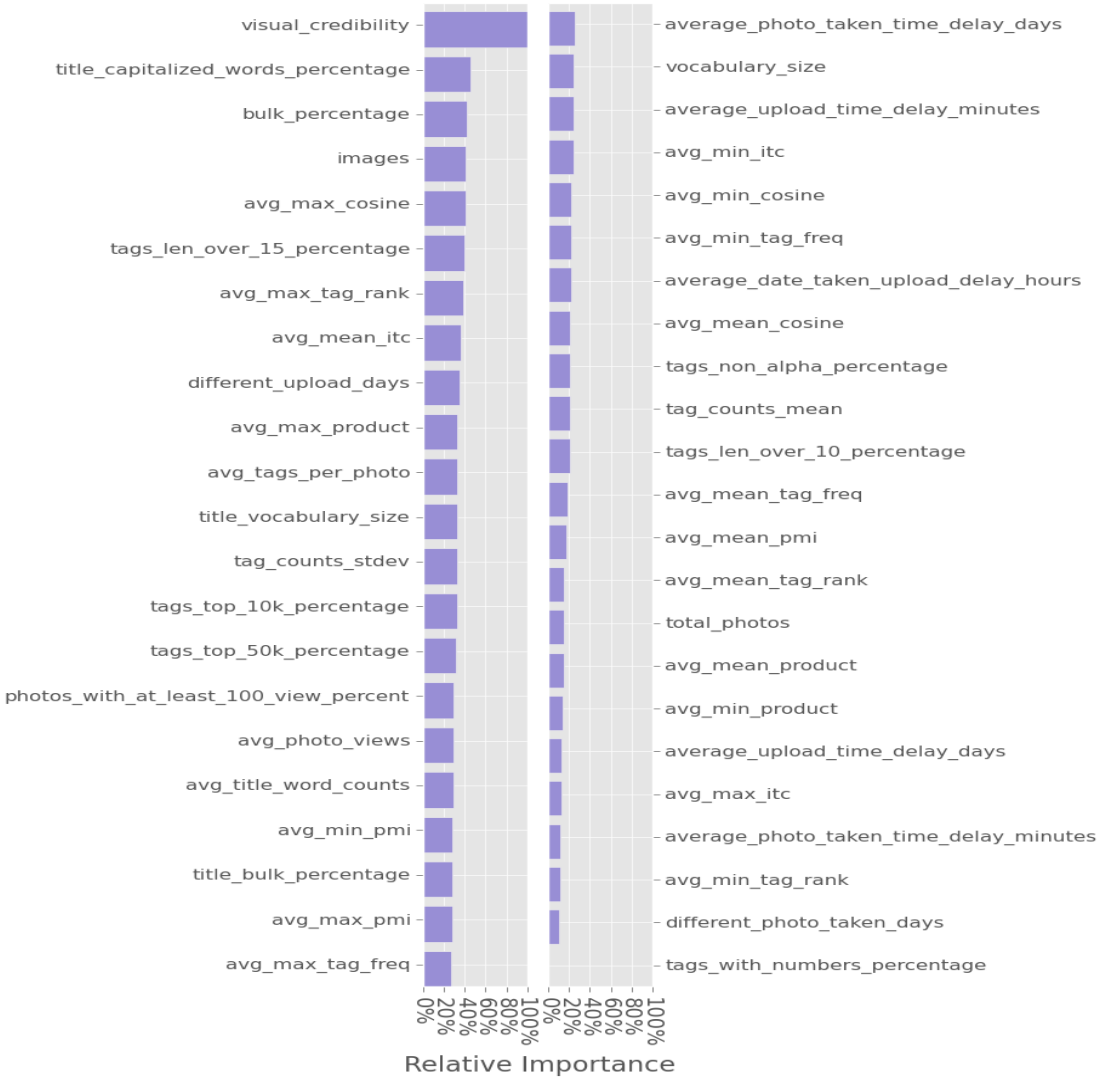


FIGURE 5.11: Features ranked according to the feature importance score provided by the GBR model on *Div150Cred*.

Table 5.2. Although the two rankings share some similarities, we also observe some important differences. We first notice that in the GBR feature importance ranking, the *avg_ upload_ time_ delay_ minutes* has the highest value, whereas in the Spearman ranking it was placed fourth. Also, photosets features play a lesser role in the model's decision than when directly looking at the correlation with the manual credibility scores. Similar to the results presented in Table 5.2, contacts features prove to be less relevant for estimating credibility.

When comparing the two proposed feature ranking methods on *MTTCred* (Figures 5.9 and 5.10), surprisingly, only the GBR method finds the *visual_ credibility* feature as the most informative one. The LR method puts the *avg_ contacts_ authority* descriptor on top, while *visual_ credibility* is ranked seventh. This observation may serve as a clue for

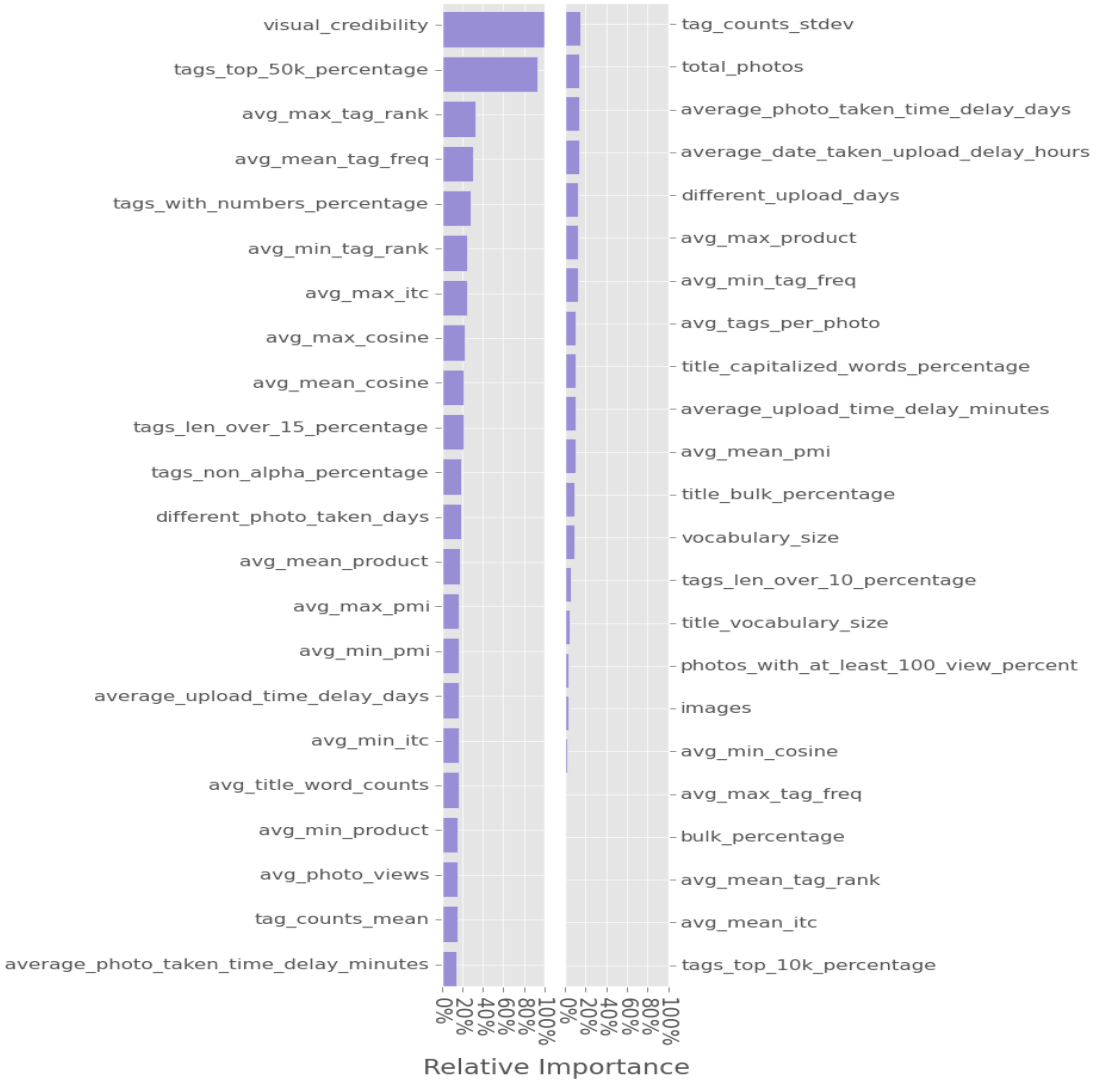


FIGURE 5.12: Features ranked according to the feature importance score provided by the LR model on *Div150Cred*.

why the GBR model outperforms LR when they are both trained on all features (Table 5.5.

If we now analyze the rankings proposed by the two methods on *Div150Cred* (Figures 5.11 and 5.12), here the *visual_ credibility* feature is placed first by both methods. However, the relative feature importance score between the first two ranked features is much higher in the GBR ranking than the LR ranking (an over 50% difference for GBR as opposed to under 10% for LR).

5.7.3 Feature selection influence

In the previous Subsection, we proposed two feature ranking methods which we applied on both on the *MTTCred* and *Div150Cred* datasets. Here, we investigate the usefulness of these rankings as feature selection methods when training regression models for predicting a user credibility score. As we can see in Table 5.9, when using all of the 45 proposed credibility estimates lowers the correlation of the best predicted credibility score, as opposed to the configuration when we train using only content features. This suggests that some features have a negative impact on regression model training. Feature selection is a possible solution for this problem.

We chose to test on both evaluation datasets using two models: Logistic Regression (LR) and Gradient Boosting Regressor (GBR) in order to have a representative from each class of regression models (*i.e.* linear and ensemble). For each model, we compare 3 feature ranking methods: the GBR and LR rankings introduced in the previous Subsection and the ranking provided by the Spearman correlation score with the ground truth credibility score. This leads to the 6 configurations noted in the legends of Figures 5.13, 5.14 and 5.15. For each configuration, we train models using top k ranked features where k ranges from 1 to the total number of evaluated features, depending on the test collection and the feature family. Each model is evaluated in a leave-one-out-cross-validation (LOOCV) fashion.

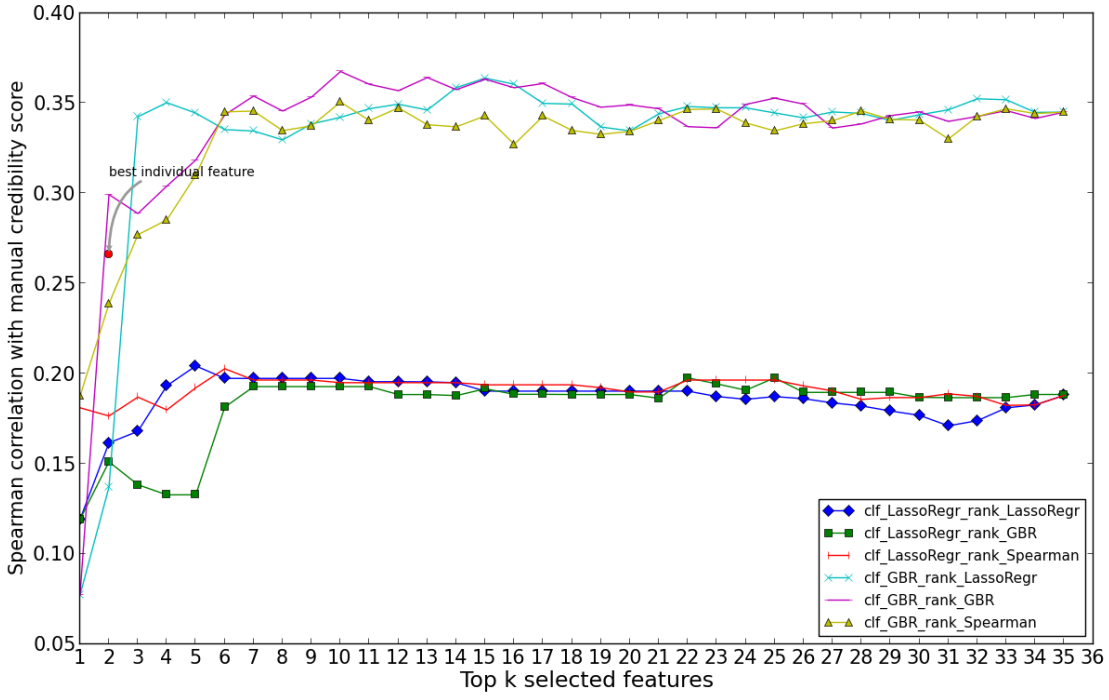


FIGURE 5.13: Impact of context features ranking methods on model learning. We test on the *MTTCred* dataset.

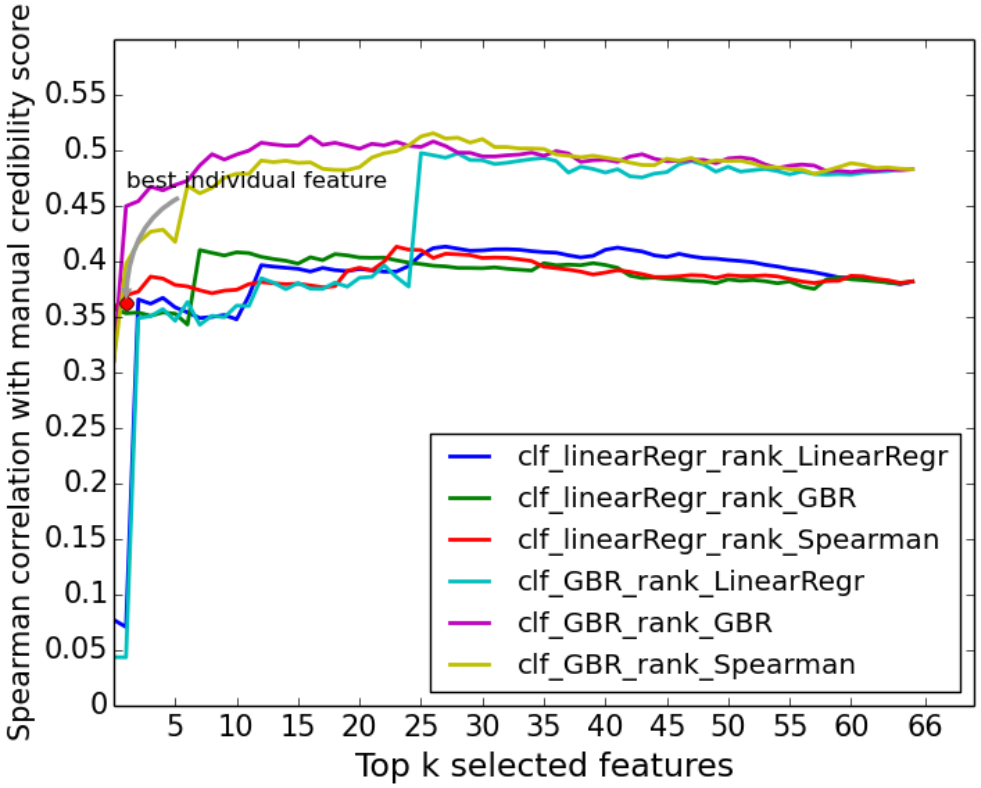


FIGURE 5.14: Impact of context + content features ranking methods on model learning. We test on the *MTTCred* dataset.

In Figure 5.13 we evaluate the impact of the six proposed learning configurations using context features on the *MTTCred* dataset. This entails that we test with k ranging from 1 to 36 (the complete set of context features), leading to 216 trained regression models. We first observe that ensemble methods are better than the linear ones at almost any feature cut-off point. Also, as expected, the GBR model suffers from higher variability. An interesting result is that we can surpass the best individual feature (*photosets_avg_comments*) by training a GBR model with only a couple of features (the first two features according to the Spearman ranking or the first three features according the the GBR or Lasso rankings). Feature selection also helps improving the best Spearman score obtained by a classifier trained on all features (0.345). There are several configurations that score higher and the best one achieves a correlation score of **0.367**. This is obtained by a GBR model trained on the top 10 features ranked according to the GBR feature importance scores. This final configuration gives a 6.37% relative improvement over the best learned feature and a 37.96% relative improvement over the best single context feature.

In Figure 5.14 we evaluate the impact of the six proposed learning configurations using context and content features on the *MTTCred* dataset. This entails that we test

with k ranging from 1 to 66 (the complete set of context and content features), leading to 396 trained regression models. Similar to the results presented in Figure 5.13, ensemble methods are better than the linear ones at almost any feature cut-off point and using less than 5 features for learning, we already surpass the best individual feature (*visual_ credibility*).

In Figure 5.15 we evaluate the impact of the six proposed learning configurations using context and content features on the *Div150Cred* dataset. This means that we test with k ranging from 1 to 45 (the complete set of context and content features), leading to 270 trained regression models. Here, as opposed to the results reported for *MTTCred*, there is little difference between the GBR models and the LR models. However, the benefits of feature selection are still quickly noticeable. For most configurations, using only 3 features for training allows us to learn a model capable of predicting a credibility estimate that is better correlated with the ground truth scores than the best individual feature (*visual_ credibility*). For example, as we can see in Table 5.11 training a GBR model on the top 3 ranked features according to the LR coefficients, we obtain a 21.62% relative improvement.

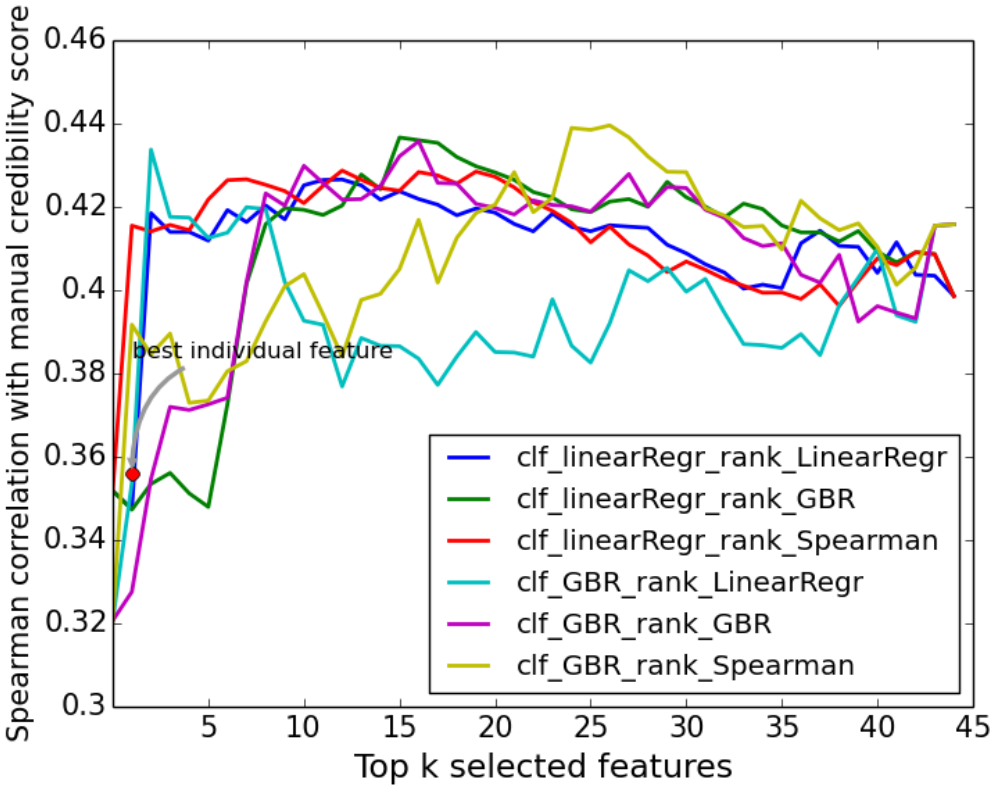


FIGURE 5.15: Impact of context + content features ranking methods on model learning. We test on the *Div150Cred* dataset.

Model	**Ranking method**	**Max Spearman**	**# of kept features**
LR	Spearman	0.413	24
	LR_coef	0.413	28
	GBR	0.410	**8**
GBR	Spearman	**0.515**	27
	LR_coef	0.497	26
	GBR	0.512	17

TABLE 5.10: Best correlation score and the number of retained features for each feature ranking and model configuration on *MTTCred*.

In Table 5.10, we give a summary of the results illustrated in Figure 5.14. We provide the best correlation scores and the number of retained features for each feature ranking and model configuration on *MTTCred*. We first notice a consistent advantage of GBR models over LR based training configurations (0.515 correlation for GBR with Spearman feature ranking vs. 0.413 for LR with Spearman or LR feature ranking). However, when comparing the models trained on the GBR ranked features, the LR model reaches its peak score when using only 8 features for training, opposed to 17 in the case of the GBR model. The best configuration (GBR model trained with the first 27 features ranked according to the Spearman method) gives a 6.62% relative improvement over the best model trained on the complete set of 66 features and a 42.26% relative improvement over the best individual feature (*visual_credibility*).

Model	**Ranking method**	**Max Spearman**	**# of kept features**
LR	Spearman	0.428	13
	LR_coef	0.426	13
	GBR	0.436	16
GBR	Spearman	**0.439**	27
	LR_coef	0.433	**3**
	GBR	0.435	17

TABLE 5.11: Best correlation score and the number of retained features for each feature ranking and model configuration on *Div150Cred*.

In Table 5.10, we give a summary of the results illustrated in Figure 5.15. We provide the best correlation scores and the number of retained features for each feature ranking and model configuration on *Div150Cred*. Compared to the results on *MTTCred*, here, there is little difference between GBR and LR based training configurations. The best

configuration is the same as for *MTTCred* (Table 5.10), GBR model trained with the first 27 features ranked according to the Spearman method. It offers a 5.2% relative improvement over the best model trained on the complete set of 45 features and a 23.31% relative improvement over the best individual feature (*visual_ credibility*).

In summary, we confirmed that using feature selection improves the regression model prediction and noticed that this improvement is more noticeable on *MTTCred* than *Div150Cred*.

5.8 Conclusion

In this Chapter, we first introduced a new dataset specifically built to help us evaluate potential indicators for credibility but also to serve as a training dataset on which we can compare multiple learning models and features. We presented the motivation behind the need for such a dataset and our methodology used for the creation of the dataset.

In the following Sections, we investigated the use of context and content features in estimating the credibility of Flickr users. We mined the features from various data sources in which a Flickr user has contributions, such as Flickr groups, photo favorites, a user's photosets or a user's contacts network. For the set of extracted features, we described the data acquisition process and tested their usefulness as individual credibility estimators. We also defined a credibility prediction problem, in which we learn regression models that provide better credibility estimators than the individual features. We find that, although individual context features are weak indicators for credibility, by choosing the appropriate regression model and the right set of features for training we are able to predict a credibility score that has considerably better correlation to the manual credibility score than any of the individual features.

Chapter 6

Practical uses of user credibility estimators

This Chapter illustrates the use of credibility estimates in two different scenarios. Firstly, user credibility estimates are embedded into a diversified image retrieval framework. Our approach is validated on a publicly available dataset (DIV400) Results indicate that a reranking of retrieved images based on user credibility is beneficial for performance. We then showcase the use of the MTTCred dataset and of its associated credibility features introduced in the previous chapter in two scenarios, i.e. user credibility classification and credible user retrieval. We find that using off-the-shelf learning models allows the exploitation of the proposed features to accurately differentiate between credible and non-credible users and to provide a relevant user ranking.

6.1 Motivation

In the previous Chapter, we introduced a new evaluation collection (*MTTCred*) and a large set of user credibility estimates. The proposed an in-depth analysis of user credibility on the Flickr platform is carried both on *MTTCred* and on *DIV150Cred*, a publicly available dataset. While this contributes to research about Web data quality in the multimedia domain, we are also interested in studying the usefulness of user credibility estimates in practical scenarios such as social image retrieval and expert identification.

User credibility for image retrieval results diversification. Existing image retrieval systems exploit textual or/and visual information to return results. Retrieval is mostly focused on data themselves and disregards the users. In Web 2.0 platforms, the quality of annotations provided by different users can vary strongly. To account for this variability, we complement existing methods by introducing user tagging credibility in the retrieval process.

Automatic credibility estimation is a recent trend in Web content analysis; it is mostly applied to textual documents, such as tweets [35] or Web pages [36]. Also related is the automatic assessment of crowdsourcer credibility, which is investigated in [37]. However, none of these works is focused on multimedia content and literature regarding multimedia credibility is limited. Xu et al. [38] aim to help users filter multimedia news by targeting credible content. They propose methods to evaluate multimedia news by comparing visual descriptions and textual descriptions respectively, as well as their combination. Yamamoto and Tanaka [39] have built ImageAlert, a system that focuses on text-image credibility. While interesting, existing work on multimedia content credibility estimation is preliminary and deserves further investigation. The estimation of individual tag relevance is related to our work. Li et al. [370] have proposed a neighbor voting framework which exploits neighbor voting to assess tag quality. More recently, Gao et al. [371] introduce a hypergraph framework to jointly model visual and textual cues of social media images. Their approach compares favorably to other existing methods but has a high computational cost at query time. We estimate credibility independently of a given topic and thus drastically reduce processing complexity at query time. [370, 371] do not aggregate relevance at user level and focus on individual tags. Both works need a large amount of data annotated with targeted tags and their efficiency on less common tags is questionable.

User credibility for expert identification. Features extracted from a user's activity in the community have been successfully used to classify users in several social media platforms. In [372], the authors use, among other indicators, statistics about the user's immediate network (e.g., number of followers/friends) and communication behavior (e.g., retweet frequency) to classify latent attributes Twitter users, including gender, age or regional origin. A combination of features extracted both from the user's profile and interactions in the community and from user generated content have been proposed for expert identification in community question answering websites. Liu et al.[161] use a vector space model to represent the question and user profiles as term vectors. The proposed expert-finding method compares the similarity of questions and user profiles and takes into consideration the differences of expertise level, posting time of query, and the number of replies to questions. In [373], the authors propose an approach that

considers user subject relevance, user reputation and authority of a category in finding experts. There, a user's subject relevance is defined as the relevance of his or her domain knowledge to the target question, user's reputation is derived from the user's historical question-answering records, while user authority is derived from link analysis. In [312] and [374], the authors focus on temporal cues that contribute to expert identification and discuss their influence in community dynamics. One important reported finding is that the temporal cues based method outperforms user statistics based ones. These works are all focused on textual content and, to our knowledge, there is no prior work on multimedia expert retrieval Inspired from the previously mentioned works, in the second part of this Chapter we propose an adaptation of expert classification and retrieval scenarios. We redefine these tasks and adapt them for credible user identification in the multimedia domain.

6.2 Improving diversity in a image retrieval system with user credibility

6.2.1 Problem definition

Existing works have identified relevance and diversity as two core properties of efficient image retrieval systems. Given that these two characteristics are antinomic, different methods have been proposed to find a good compromise between them. Classically, relevance was primarily estimated by using textual weighting schemes. However, with the improvement of low-level image descriptors, multimedia fusion schemes also gained traction. Diversity is usually improved by applying clustering algorithms which rely on textual or/and visual cues [375]. In addition, the usefulness of social cues was also explored for Web 2.0 platforms [339] but this aspect remains secondary.

Our work is focused on the estimation and exploitation of user credibility, a cue which was not previously exploited in multimedia retrieval and is complementary to those cited above. Here, we investigate user tagging credibility in the context of image retrieval result diversification and we focus on the following questions:

- Q_1 - how should credibility be integrated in existing multimedia retrieval systems?
- Q_2 - what is the additional complexity of credibility estimation?

We test the usefulness of user credibility estimates in the setting proposed by the *Retrieving diverse social images* MediaEval benchmark initiative [376]. The main objective of the evaluation from [376] is to maximize result diversity, which is captured with cluster

TABLE 6.1: Statistics of the DIV400 dataset.

	Devset	**Testset**
#topics	50	346
#images	5,118	38,300
#users	1,154	5,362

recall at N (CR@N), that accounts for the number of different clusters represented in the top N results. However, since a good retrieval method should find a good compromise between relevance and diversity, we also two other usual retrieval metrics. P@N performance counts the number of relevant images in the top N results without considering clusters and is thus relevance oriented. Finally, the F1@N, the harmonic mean of CR@N and P@N, is also used to evaluate the combination of diversity and relevance. CR@10 was the main official metric in [376] but we also report results at 20 and 30 recall depths.

6.2.2 Dataset

We evaluate our retrieval method with the *DIV400* dataset, which is thoroughly described in [376] and is summarized in Table 6.1. The topics represent tourist points of interest (POIs). It consists of a development dataset (50 tourist POIs, 5,118 photos) and a testing dataset (346 POIs, 38,300 photos). Each POI is represented with up to 150 photos and associated metadata retrieved with Flickr's default "relevance" algorithm. Data is collected with both textual and GPS queries and also includes images and a wide range of POI metadata. Relevance and diversity annotations are available for each photo. Photos are considered relevant if they depict a common photo representation of the POI. A set of photos is considered to be diverse if it depicts complementary visual characteristics of the target POI. Clusters are manually built from relevant images of each POI.

6.2.3 Dataset Processing

Our diversification approach is mainly based on visual content mining. To keep abreast with recent advances in computer vision, we use Overfeat [14], a powerful CNN-based feature, to model user credibility and to process the *DIV400* images. PCA is applied to these features to obtain a more compact representation of images and thus accelerate retrieval. Preliminary tests have already shown that results obtained with the first 256 PCA dimensions are equivalent to those obtained with the default Overfeat configuration (4096 dimensions).

Inspired by [32], face and blur detection are applied to remove images with salient faces and blurred images, which are potentially irrelevant for a part of the topics. Face detection is implemented with the standard OpenCV algorithm[1]. Preliminary tests has shown that direct removal of images containing faces does not improve results. Consequently, given a set of POI images and the associated user set t_u, face removal is performed based on p_u, the proportion of users from the set t_u which upload face images. Face images are retained for p_u values lower than a threshold ($th(p_u)$) and discarded otherwise. In order for p_u to be meaningful, we impose face removal only on the POIs with at least $th(t_u)$ contributors. p_u exploits social consensus about usefulness of face images and is optimized on the devset of *DIV400*. Blur detection is performed using thresholded gradient. Similar to face retrieval, a threshold $th(b)$ for blur removal is learned on the devset of DIV400.

6.2.4 Proposed approach

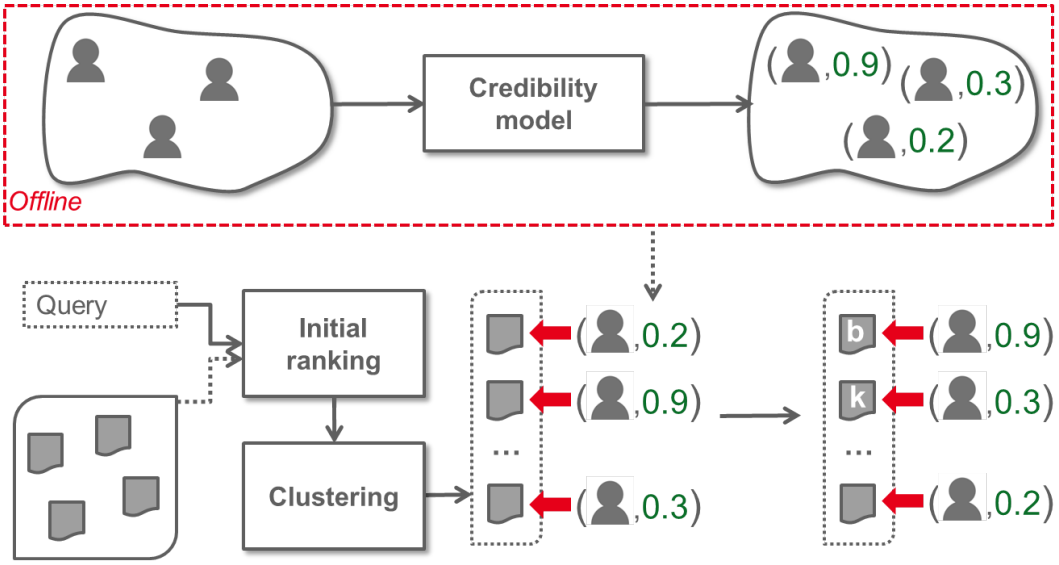


FIGURE 6.1: Using credibility estimations for diversification.

We propose a retrieval method which diversifies images using k-Means clustering and improves relevance with credibility estimations. In the framework detailed in Figure 6.1, the credibility estimate of a user can be any individual credibility feature or learned feature introduced in Chapter 5. In this Section, our goal is to provide a proof-of-concept on the usefulness of user credibility in an image diversification scenario. Also, given the large number of users that have contributions in the *DIV400* dataset (Table 6.1), we chose to test our credibility based retrieval framework only with the *visual_credibility* feature as a credibility estimator. This choice is motivated by the fact that this feature

[1] http://opencv.org/

is the best performing individual feature on two user credibility evaluation datasets (see 5). A detailed presentation of the extraction process of this feature is presented in Section 5.5 of Chapter 5. Given that the *DIV400* and *DIV150Cred* datasets share the same underlying domain, we chose the configuration of the *visual_ credibility* feature according to the results described in Table 5.4. There, for *DIV150Cred*, the best Spearman correlation with the ground truth credibility score was obtained averaging the predictions of ImageNet based models (see Chapter 3).

Clustering is performed using the L2-normalized version of the features obtained with the default configuration of Overfeat [14]. Let $\mathbf{L_F} = \{(I_1, U_1), (I_2, U_2), (I_3, U_1), ..., (I_N, U_M)\}$ be the ranked list of Flickr images which should be reranked. Here (I_i, U_j) denote image-user pairs. Our retrieval method can be broken down into three steps: initial filtering, cluster ranking and image sorting.

Initial filtering In this step, we simply remove from $\mathbf{L_F}$ all pairs (I_i, U_j) for which I_i qualifies for face or blur removal.

Cluster ranking After image filtering, we perform k-Means clustering to diversify the topic representation. Let $\mathbf{C_F} = \{C_1, C_2, ..., C_k\}$ be the clustered version of $\mathbf{L_F}$. Inspired by [339], we rank clusters based on $\#Users$, the number of distinct users which contribute to each cluster. Ranking based on $\#Users$ gives priority to clusters which show social consensus. When ties appear with $\#Users$, they are broken by using the the user with the highest credibility score $cred(U)$ from each cluster. As a result, we obtain $\mathbf{C_F^R} = \{C_3, C_k, C_2, ..., C_1\}$, a list of clusters ranked using social cues. For comparison, we also rank clusters based on their raw image count ($\#Images$).

Image sorting We exploit credibility estimation to sort images within clusters. Let $C_c = \{(I_1, U_1), (I_3, U_5), (I_8, U_1)\}$ be a cluster with its images ranked by Flickr. Assuming that $cred(U_5) > cred(U_1)$, the sorted representation of the cluster will be $C_c^R = \{(I_3, U_5), (I_1, U_1), (I_8, U_1)\}$. In C_c^R, the sorted version of C_c priority is given to images uploaded by users with higher credibility score. The final image ranking $\mathbf{L_F^R}$ is obtained by iterating over $\mathbf{C_F^R}$, the ranked list of clusters, and by selecting each time the first unseen image from C_c^R.

6.2.5 Experiments and results

6.2.5.1 Clustering Analysis

Choosing the number of clusters in a diversification task is a hard problem and we experiment with different values of this parameter. In Figure 6.2, we illustrate the impact of the number of clusters on clustering performances. Cluster ranking is performed either based on the number of users which contributed to the cluster (*#Users*) or on the number of images a cluster contains (*#Images*). The *#Users* based cluster ranking outperforms *#Images* based ranking.

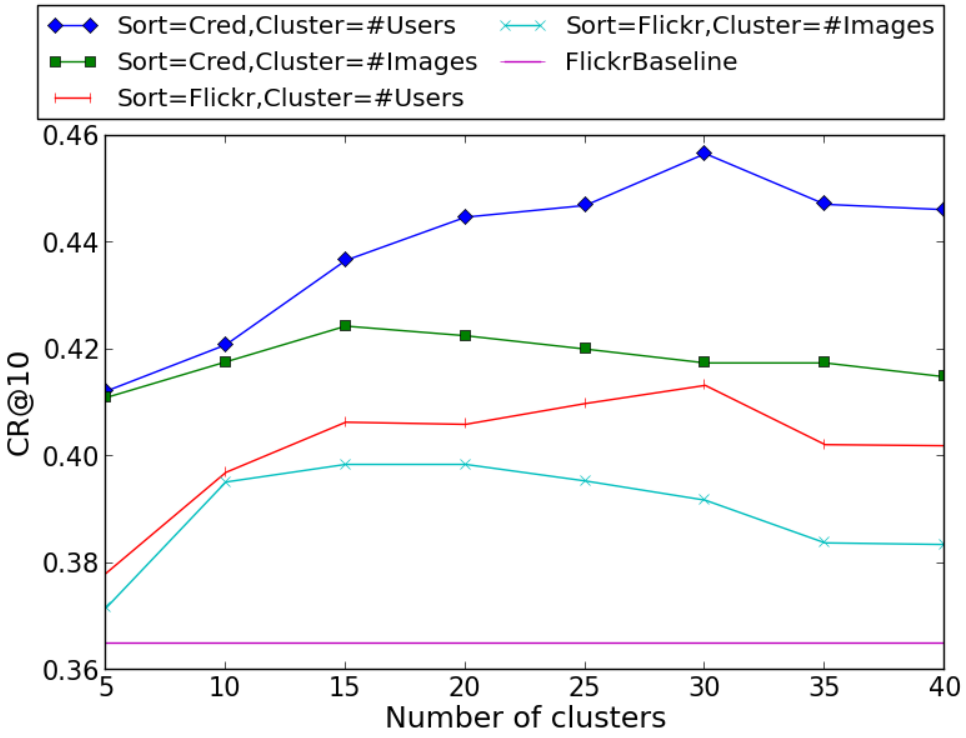


FIGURE 6.2: CR@10 performances with different clustering methods and different numbers of clusters on the testset of DIV400. *Sort* denotes the type of image sorting used within clusters. *Cred* is a sorting based on user credibility and *Flickr* is the original Flickr ordering. "Cluster" denotes the cluster ranking method. *#Users* and *#Images* represent the user and image counts of a cluster.

Within each cluster, *Cred*, the credibility based image sorting outperforms the use of the initial Flickr sorting in all settings. Intuitively, the best overall results are obtained when *#Users* and *Cred* are combined for inter- and intra-cluster ranking. With 30 clusters, *Flickr* + *#Users* brings a 2 CR@10 points improvement of results compared to *Flickr* + *Images*. This result confirms the conclusions of [339], namely that the use of social cues for cluster ranking is beneficial. More importantly, the introduction of credibility estimation (*Cred* + *#Users*) further improves CR@10 by 4 points. We present results

on the testset here because they are obtained by averaging a larger number of topics. However, similar results are obtained on the devset and *Cred* + *#Users* with 30 clusters is used for further experiments.

6.2.5.2 Global performances

In table 6.2, we present the results obtained with the best credibility based retrieval method, described in Section 6.2.4. It combines clustering and user credibility estimates and produces a reranked list of images $\mathbf{L_F^R}$. For comparison, we also present results obtained by the two most efficient existing methods tested on DIV400 [376].

To understand the impact of face and blur removal, we briefly present results obtained when we skip one of these steps. When no prefiltering is used CR@10 is 0.4437. The use of blur removal or of face removal augments the score to 0.4476 and to 0.4536 respectively. While image filtering is beneficial, the main contribution comes from the use of credibility and of user centered clustering.

TABLE 6.2: Comparison of retrieval results obtained with different methods on DIV400 and CR@N, P@N and F1@N metrics. SOTON-WAIS [32] and SocSense [33] are the two most efficient retrieval methods proposed at MediaEval Diverse Images 2013. $\mathbf{L_F^R}$ corresponds to a setting with *Cred*+*#Users* and 30 clusters (Figure 6.2).

Method	metrics	@10	@20	@30
SOTON-WAIS [32]	P	**0.8158**	**0.7788**	0.7414
	CR	0.4398	0.6197	0.7216
	F1	0.5455	**0.6607**	0.7019
SocSens [33]	P	0.733	0.7487	**0.7603**
	CR	0.4291	0.6314	0.7228
	F1	0.5209	0.6595	**0.7087**
$\mathbf{L_F^R}$	P	0.7822	0.7154	0.6927
	CR	**0.4567**	**0.6582**	**0.7801**
	F1	**0.5526**	0.659	0.7073

A comparison of our method to [32] and [33] shows that cluster recall is improved at all cut-off points. For CR@10, the official metric associated to DIV400, the improvement is close to 2 and 3 points respectively. Confirming other results obtained on DIV400, which show that clustering hurts precision, the P@10 obtained with $\mathbf{L_F^R}$ is lower than those obtained in [32]. However, the F1@10 score of our method is slightly better and this comparison shows that our approach is competitive. It also departs from existing retrieval methods by the important role given to social cues and particularly to credibility.

6.3 User credibility for expert retrieval

In this Section, we propose two novel use cases for user credibility estimates. We investigate the relevance of the credibility features introduced in the previous Chapter both on a classical multi-class supervised learning problem adapted to user credibility and a retrieval task inspired by expert retrieval works, in which we rank users based on predicted credibility scores. We are also interested in showcasing the use of the newly developed user tagging credibility dataset, *MTTCred* (see Chapter 5 for dataset details), on these two tasks. While the focus is on evaluating on our proposed test collection, we also use the domain specific *Div150Cred* credibility dataset throughout this Section.

6.3.1 Problem Definition

Like most of the works that deal with predicting credibility in social media, such as the credibility of tweets [258], the problem is viewed as a classification problem. In those scenarios, two (credible / not credible) or several credibility classes are considered. Here, we first define a classification problem in which we have 5 credibility classes depending of the users' ground truth credibility score as follows:

- *C1:* highly not credible users - credibility score $\in [0, 0.2)$.
- *C2:* not credible users - credibility score $\in [2, 0.4)$.
- *C3:* uncertain credibility - credibility score $\in [0.4, 6)$.
- *C4:* credible users - credibility score $\in [0.6, 0.8)$.
- *C5:* highly credible users - credibility score $\in [0.8, 1]$.

TABLE 6.3: Distribution of users by class for the *MTTCred* and *Div150Cred* datasets.

User classes distribution	**Dataset**	
	MTTCred	*DIV150Cred*
# of users in C1	67	55
# of users in C2	394	69
# of users in C3	424	199
# of users in C4	117	281
# of users in C5	7	81

For the credible user retrieval task, we propose different unions of the credibility classes, dictated by the different evaluation measures that we use. We will provide more details in Section 6.3.5.

6.3.2 Credibility features

We gained insight on which are the most relevant features for predicting a user's credibility score in Section 5.7.3 of Chapter 5. Looking at the overview results presented in Table 5.10, we chose to use the first 17 features ranked by the GBR method (see Figure 5.9) for the experiments carried on *MTTCred*. We note this subset as $Features_{MTTCred\_selected}$. Similarly, inspired by the results presented in Table 5.11, we use the first 17 features ranked by the GBR method (see Figure 5.11) for the experiments carried in this Section on *DIV150Cred*. We note this subset as $Features_{Div150Cred\_selected}$.

Looking at the overview results presented in Table 5.11, we chose to Similarly,

In both cases, the GBR ranking is slightly less efficient than the Spearman one (0.003 difference of Spearman correlation on *MTTCred* and 0.004 on *Div150Cred*. We chose to use the features proposed by the GBR ranking given that we obtain these scores with only 17 features, opposed to 27 in the case of the Spearman ranking.

For the *Div150Cred* dataset, we also experiment with an extended set of features (noted $Features_{Div150Cred\_selected+domain}$ introduced in [290] that are specific for user tagging credibility in the landmark photo retrieval domain. Next, we list these features, as described by Ionescu et al. [290]:

- *face_proportion*: Feature obtained using the same set of images as for the *visual_credibility* feature introduced in Chapter 5. The default face detector from OpenCV[2] is used here to detect faces. *face_proportion*, the percentage of images with faces out of the total of images tested for each user is computed. The intuition behind this descriptor is that the lower *face_proportion* is, the better the average relevance of a user's photos is. *face_proportion* is normalized between 0 and 1, with 0 standing for no face images.
- *tag_specificity*: Feature obtained by computing the average specificity of a user's tags. Tag specificity is calculated as the percentage of users having annotated with that tag in a large Flickr corpus (≈100 million image metadata from 120,000 users).

[2]http://docs.opencv.org/trunk/doc/py_tutorials/py_objdetect/py_face_detection/py_face_detection.html

- *location_similarity*: Feature obtained by computing the average similarity between a user's geotagged photos and a probabilistic model of a surrounding cell of approximately $1km^2$ geotagged images. These models were created using the model in [377]. The intuition here is that the higher the coherence between a user's tags and those provided by the community is, the more relevant her images are likely to be.

To compare these features with the credibility descriptors introduces in Chapter 5, we list the Pearson correlation value with the ground truth credibility score reported in [290]: -0.2687 for *face_proportion*, -0.2883 for *tag_specificity* and 0.1329 for *location_similarity*. Although we use the Spearman correlation in in Table 5.6 and the values are not directly comparable for the features, the high correlation scores for the features presented above make them good candidates for our credible user classification and retrieval experiments on *Div150Cred*.

6.3.3 Data exploration

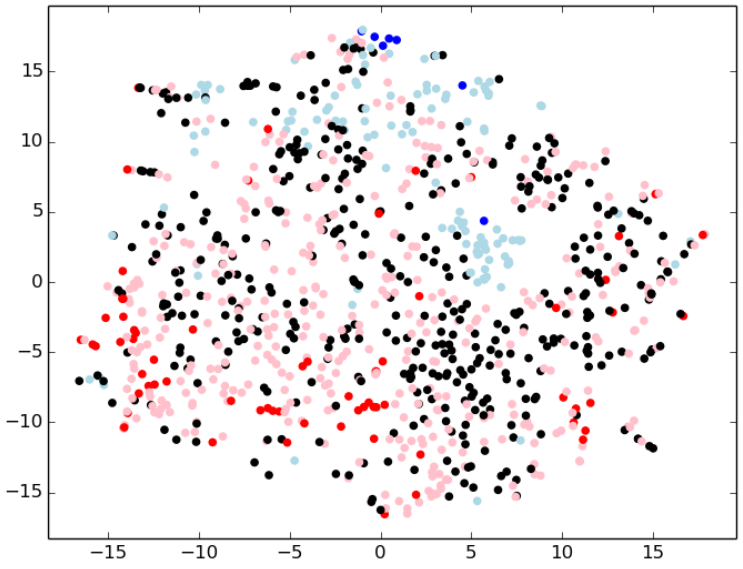


FIGURE 6.3: Visualization of the 1009 users from the *MTTCred* dataset using the t-SNE algorithm. The values from both axes are automatically determined by t-SNE. The strong blue points represent users from the *C5* class, pale blue the ones from the *C4* class, while strong red and pale red represent users from the *C1* and *C2* classes, respectively. Black points correspond to *C3* users.

Each user from the *MTTCred* dataset is described by 17 features. In Figure 6.3, we provide a visualization of a projection of those features in the two dimensional space for the 1009 users using the t-SNE algorithm [378]. We first observe in the upper left corner

a small cluster including 4 out of the 7 highly credible users. On the contrary, towards the right side of the plot we can see the users belonging to the *C1* class. Although most of the users fall under the *uncertain credibility* category and are scattered all over the plot, the dotted black lines mark a separation between most of the credible users and the others. Just by looking at this plot, we can assume that a non linear classifier can potentially be able to discern between credible and non credible users. We will show in the next section that this hypothesis is partially confirmed.

6.3.4 User classification experiments

Given the fact that we have few instances in our datasets (1009 users in *MTTCred* and 685 in *Div150Cred*), we afford to perform tests using a Leave-One-Out Cross Validation (LOOCV) method. For example, in the case of *MTTCred*, on each iteration, we train a model on 1008 users and predict for the one left aside. Before the classification, all the features are L2 normalized. We tested several classifiers and the best accuracy scores, reported in Tables 6.4, 6.5 and 6.6 are obtained with an Extra Trees Classifier model. For all the experiments in this Section, we perform parameter tuning and compare models from the scikit-learn toolkit [379].

TABLE 6.4: Confusion matrix of user credibility class prediction on *MTTCred* using the $Features_{MTTCred\_selected}$ feature set.

Predicted Class							**Accuracy**
True Class		**C1**	**C2**	**C3**	**C4**	**C5**	
	C1	51	15	1	0	0	0.761
	C2	3	275	114	2	0	0.697
	C3	0	101	312	11	0	0.735
	C4	0	9	32	75	1	0.641
	C5	0	0	1	4	2	0.296
Overall Accuracy							**0.708**

Considering that the main goal of this experiment is to analyze the potential of multi-class classification on our proposed dataset and not to maximize the accuracy score, we

TABLE 6.5: Confusion matrix of user credibility class prediction on *Div150Cred* using the $Features_{Div150Cred\_selected}$ feature set.

Predicted Class							**Accuracy**
True Class		**C1**	**C2**	**C3**	**C4**	**C5**	
	C1	19	25	11	0	0	0.345
	C2	3	24	18	24	0	0.347
	C3	2	5	120	69	3	0.603
	C4	1	0	33	217	30	0.772
	C5	0	2	6	53	20	0.246
Overall Accuracy							**0.583**

TABLE 6.6: Confusion matrix of user credibility class prediction on *Div150Cred* using the $Features_{Div150Cred\_selected+domain}$ feature set.

Predicted Class							**Accuracy**
True Class		**C1**	**C2**	**C3**	**C4**	**C5**	
	C1	26	20	9	0	0	0.472
	C2	2	36	15	16	0	0.521
	C3	2	4	133	58	2	0.668
	C4	1	0	23	237	20	0.843
	C5	0	2	5	50	24	0.296
Overall Accuracy							**0.665**

still obtain a good overall accuracy (0.692) using only 17 credibility features. While proposing a fine-grained user classification task renders the classification problem more difficult, it allows us to dwell in a deeper analysis of Flickr user credibility. Although the accuracy scores for individual classes are not very high, the confusion matrix presented in Table 6.4 gives us an insight on where the classifier makes mistakes. As it can be also observed in Figure 6.3, most of the misclassifications fall in the *C3* class. We also consider a possible real world scenario, similar to tweet credibility classification, where we are interested to differentiate between credible and non credible users and disregard the degree of credibility and the users of *uncertain credibility*. This entails that we will have a *Cred* class composed by the union of *C4* and *C5* and a *NotCred* class, containing users from *C1* and *C2*. In this case, we obtain a 0.888 accuracy for the *Cred* class and 0.994 for the *NotCred* one.

In Table 6.5, we present the confusion matrix of user credibility class prediction on *Div150Cred* using the $Features_{Div150Cred\_selected}$ feature set, while in Table 6.6, we show the confusion matrix of user credibility class prediction on *Div150Cred* using the $Features_{Div150Cred\_selected+domain}$ feature set. When comparing the global accuracy results from Table 6.5 to those from Table 6.6, we can observe a 14.06% relative increase of the accuracy when adding domain specific features.

6.3.5 Credible users retrieval experiments

In this Section, we describe how the *MTTCred* and *Div150Cred* datasets can be used for a credible user retrieval task. In order to obtain a user ranking, we employ a LOOCV method but unlike the models used in the previous section, we test regression models that predict a credibility score instead of the credibility class. The users are ranked in descending order of the predicted credibility scores and the comparison is done between the manual ranking and that obtained with the different tested methods.

TABLE 6.8: Comparison of regression models for credible user retrieval on *Div150Cred* using the $Features_{Div150Cred\_selected}$ feature set.

Model	**Metric**					
	P@10	*P@100*	*AP*	*NDCG@10*	*NDCG@100*	*NDCG*
LR	0.4	0.3	0.217	0.213	0.521	0.762
SVR	0.4	0.36	0.328	0.239	0.53	0.765
ETR	0.5	0.52	0.507	**0.582**	**0.692**	**0.81**
RFR	**0.6**	0.53	0.519	0.56	0.675	0.892
GBR	**0.6**	**0.58**	**0.542**	0.568	0.678	0.804

TABLE 6.9: Comparison of regression models for credible user retrieval on *Div150Cred* using the $Features_{Div150Cred\_selected+domain}$ feature set.

Model	**Metric**					
	P@10	*P@100*	*AP*	*NDCG@10*	*NDCG@100*	*NDCG*
LR	0.5	0.59	0.502	0.288	0.582	0.803
SVR	0.5	0.61	0.517	0.307	0.579	0.816
ETR	0.6	0.67	0.64	**0.648**	**0.749**	**0.924**
RFR	0.7	0.66	0.635	0.637	0.713	0.908
GBR	**0.8**	**0.69**	**0.672**	0.642	0.736	0.917

TABLE 6.7: Comparison of regression models for credible user retrieval on *MTTCred* using the $Features_{MTTCred\_selected}$ feature set.

Model	**Metric**					
	P@10	*P@100*	*AP*	*NDCG@10*	*NDCG@100*	*NDCG*
LR	0.2	0.42	0.368	0.193	0.532	0.853
SVR	0.2	0.43	0.362	0.236	0.548	0.861
ETR	**0.8**	0.58	0.551	0.602	0.728	0.912
RFR	0.7	0.59	0.516	0.442	0.682	0.896
GBR	0.7	**0.61**	**0.594**	**0.615**	**0.731**	**0.918**

In Tables 6.7, 6.8 and 6.9 we compare a set of regression models with several standard metrics used to test the relevance of ranked lists in regards to ground truth labelings. We test both linear models, such as Linear Regression and Support Vector Regression (SVR) and ensemble models, such as Extra Trees, Random Forests (RF) and Gradient Boosting (GB) Regressors. Similar to the evaluation protocol for expert retrieval in social networks described in [380], we consider the following metrics: Precision at two cut-off points (10 and 100), Average Precision over the complete list (AP), Normalized Discounted Cumulative Gain (NDCG) at 10, 100 and for the full list. While AP provides a compact measure of the precision of the retrieval capability, NDGC measures the ability of a model to retrieve different levels of credible users at high positions in the result set. P@10 and NDCG@10 are well suited for understanding the perceived quality of the first

10 retrieved users. For the precision metrics, we consider each user with a ground truth score higher than 0.6 as credible (relevant in terms of information retrieval) and the rest as not credible (not relevant). For the NDCG metrics, we consider the users from the *C1* and *C2* classes as non relevant and are given a relevance score of 0, *C3* users are given a relevance score of 1 (i.e. slightly relevant), *C4* users a score of 2 and *C5* users a score of 3. Using this approach, we can use the property of the NDCG metric of evaluation different levels of relevance in a retrieved list.

From Figure 6.3, we can see that users are scattered and is difficult to find a linear separation, even in a higher dimentional space. Confirming this observation, linear models perform poorly over all metrics. With the exception of P@100, the Extra Trees Regressor model performs the best over all other metrics. This confirms the classification results from the previous section, where the best performing model was the Extra Trees Classifier. This finding is in line with the recent findings presented in [328], in which the authors found that ensemble methods provide the best global results over a large number of diverse datasets.

For any model, we can observe an increase over the whole ensemble of retrieval evaluation metrics of $Features_{Div150Cred\_selected+domain}$ feature set over $Features_{Div150Cred\_selected}$. As in the credible user classification task, adding only three domain specific features leads to a boost in performance. For instance, we get an 20.29% relative increase of the AP. As it can be seen in Table 6.3, there are more credible users in *Div150Cred* than *MTTCred* according to our class distribution. This can explain why we get higher retrieval scores for *Div150Cred* (*e.g.* a 13.1% relative increase of AP) than for than *MTTCred*.

6.4 Conclusion

In this Chapter, we first proposed an exploration of the introduction of user tagging credibility estimation in image retrieval systems. Evaluation results show that credibility is a good complement to direct text and/or visual content analysis. Credibility estimations were integrated with a classical clustering algorithm. The performance gains obtained through the use of credibility account for its usefulness in retrieval. Finally, additional complexity is added to the retrieval framework but affects only retrieval steps which are performed offline. These steps, including feature extraction, visual model learning and credibility estimations, can be repeated periodically to follow the dataset evolution. At query time, only a reranking of images which accounts for credibility is required and this procedure has negligible effects compared to clustering.

In the second part of this Chapter, we introduced a user classification task and a credible user retrieval one. We performed tests both on the *MTTCred* dataset and on *Div150Cred* and found that ensemble models perform best on both of the proposed tasks. We also noticed that adding domain specific credibility estimates leads to better results in both scenarios.

Chapter 7

Conclusions

In this final Chapter of this Thesis, we briefly recapitulate the main contributions of our research and discuss possible directions for future work.

7.1 Summary and contributions

The main contributions of our work are essentially exposed in chapters 3 to 6.

Large scale visual concept modeling. We showed in Chapter 3 that with an appropriate choice of the initial collection and with the introduction of efficient image reranking techniques, we can train visual concept models from automatically built resources that can rival with those built from manually labeled resources. A good coverage of the conceptual space is obtained with an appropriate choice of the initial Web dataset. We explored the use of Flickr groups, but the pipeline presented here is easily applicable to larger datasets. We proposed a scalable classification framework based on binary linear models. Throughout the Chapter, we compared models trained on ImageNet data (≈ 13 *million* images covering 17 462 concepts) with models learned from Flickr groups (≈ 11 *million* images covering $38,500$ concepts). For noisy Web images, we first propose a solution in order to eliminate less visually salient groups and then compare three image ranking methods in order to select positive training instances. Finally, we obtain a collection of 30 000 Flickr groups. Taking the average cross-validation (CV) scores of the first 30,000 groups, we noticed a 2% increase, when compared to the full set of groups models. We then evaluated models by their CV training accuracy and found that models trained from Web images present a similar performance (0.924 mean CV) to those obtained from ImgeNet (0.937 mean CV). We also investigated the impact of

the image descriptors and the number of negative instances used for test on the classification process. We found that the VGG feature [65] outperformed other CNN features on all evaluation configurations. We furthermore noticed a beneficial effect of prediction accuracy when increasing the number of negatives (up to $10,000,000$).

Efficient CBIR with semantic descriptors. In Chapter 4 we proposed a technique for the automatic mining of large-scale visual resources from Web data. Based on this result, we proposed a new semantic image representation (*Semfeat*) built by aggregating individual visual concept models predictions. It was tested in content based image retrieval (CBIR) on three well known image collections (ImageCLEF Wikipedia Retrieval 2010 Collection, MIRFLICKR and NUS-WIDE). With an appropriate choice of the initial collection and with the introduction of scalable but efficient image reranking techniques, the results obtained with the automatically built resource can rival with those of the manual resource. At large scale, automatic resource construction requires significantly less effort than manual labeling and constitutes an appealing alternative to datasets such as ImageNet. Efficient semantic representations can be built through the combined use of powerful initial features (*i.e.* VGG) and of an appropriate visual representation of feature components. With the use of simple scalable learning models, as proposed in this Chapter, it is easy to scale way beyond tens of thousands of models. However, when enriching conceptual coverage, one should be careful about the potential negative effects of redundancy, a problem which deserves close attention.

We compared *Semfeat* with one of the best pre-CNN image descriptors, *Fisher Vectors* [24], three CNN image features (*Overfeat* [14], *Caffe* [81] and *VGG* [65]) and two of the best high-level image features reported in literature [5] (*PiCoDes* [4] and *Meta-class* [23]). We obtain state of the art CBIR results on the ImageCLEF Wikipedia Retrieval with several *Semfeat* configurations. The best one is given when using Flickr groups visual models trained with VGG descriptors ($MAP = 0.2127$). We get a 284.62% relative MAP improvement over Fisher and 26.38% over VGG. For comparison, the best text run submitted during the campaign, which combined annotations in different languages and sophisticated language models, had $MAP = 0.2361$ [24]. Both on MIRFLICKR and NUS-WIDE, the best retrieval results are obtained with the same *Semfeat* configuration, proving its consistency across datasets and confirming once more the usefulness of Flickr group image for visual concept learning. On MIRFLICKR, using *Semfeat* we noticed a 27.91% relative improvement of the MAP@1000 score over *VGG* and 127.4% over *Meta-class*. On NUS-WIDE, we report a 160.4% relative improvement of MAP@1000 over *VGG* and 135.7% over *Meta-class*.

In image retrieval, compactness is achieved by sparsifying semantic features (*i.e.* keeping only the top K highest individual prediction scores) and by using inverted indexes. With this scheme, very large volumes of data can be search without precision loss, as it is the case for existing dense features [25]. For 10 million images, average latency is 0.85, 13.6 and 47.8 ms for $K \in \{1, 5, 10\}$. Corresponding latencies for a collection of 100 million images are 8.6, 130.8 and 454 ms. We showed that *Semfeat* enables near real-time searches in a 1 billion images collection using a single core, provided that enough RAM is available. In our implementation, RAM consumption for 1 billion images is between 62 GB ($K = 1$) and 248 GB ($K = 4$).

In summary, the results reported in this part of our research indicate that semantic features are reliable and efficient, and finally useful for CBIR retrieval. Following [22], [23] or [3], we bring new evidence concerning the usefulness of semantic features and, in particular, propose an efficient way to clean and exploit large-scale noisy Web corpora.

User credibility in image sharing platforms. In Chapter 5, we defined the concept of *credibility* in image sharing platforms. We investigated the use of context and content features in estimating the credibility of Flickr users. We proposed and evaluated 66 user tagging credibility estimators. We essentially mined the content produced by a user—mainly tags and images. We proposed a visual credibility estimator through which it is possible to evaluate the relevance of the tags associated to a user's images using the visual concept models presented in Chapter 3. Besides these, we mined the features from various data sources in which a Flickr user has contributions—such as Flickr groups, photo favorites, user's photosets or user's contacts network. For the set of extracted features, we described the data acquisition process and tested their usefulness as individual credibility estimators. We also defined a credibility prediction problem, in which we learn regression models that provide better credibility estimators than the individual features. We found that, although individual context features are weak indicators for credibility, by choosing the appropriate regression model and the right set of features for training, we are able to predict a credibility score that has considerably better correlation to the manual credibility score than any of the individual features.

Besides investigating the usefulness of the proposed credibility estimates on a publicly available domain specific collection, *Div150Cred* [290], we also introduced in this Chapter a novel user tagging credibility evaluation dataset (*MTTCred*). Having described the process behind building this dataset, we provided detailed information about the annotation process, rater agreement scores and how we construct a user ground truth credibility score.

The dataset and the credibility descriptors introduced in this Chapter pave the road for future research towards user credibility estimation in image sharing platforms.

Practical uses of user credibility estimators. In Chapter 6, we studied the usefulness of user credibility estimates in two scenarios: social image retrieval and expert identification.

In the first part of this Chapter, we proposed to explore the introduction of user tagging credibility estimation in image retrieval systems. Evaluation results seem to already indicate that credibility is a good complement to direct text and/or visual content analysis. Credibility estimations were integrated with a classical clustering algorithm. The performance gains obtained through the use of credibility account for its usefulness in retrieval. Finally, additional complexity is added to the retrieval framework; it affects retrieval steps which are performed offline. These steps, including feature extraction, visual model learning and credibility estimations, can be repeated periodically to follow the dataset evolution. At query time, only a reranking of images which accounts for credibility is required and this procedure has negligible effects compared to clustering.

In the second part of this Chapter, we introduced a user classification task and a credible user retrieval one. We performed tests both on the *MTTCred* dataset and on *Div150Cred* and found that ensemble models perform best on both of the proposed tasks. We also noticed that adding domain specific credibility estimates leads to better results in both scenarios.

7.2 Perspectives and future work

There are several ways to further improve and extend the work presented in this Thesis. Firstly, these may focus on incorporating recent ideas from computer vision research into our work. Secondly, we propose novel domains that could benefit from our multimedia credibility analysis framework and the usage of Web data for semantic image description. We briefly describe these perspectives in what follows.

Negative instances selection for visual concept learning. In Chapters 3 and 4, we evaluated the impact of the number of negative samples used for visual concept learning. These were taken from a single large negative image collection. It has been shown that adapting the negative class for each concept improves the model's performance [355]. However, this approach adds an increased level of complexity; whether it scales for the large number of concepts that we model in this Thesis remains an open question.

Concept selection and hierarchies in Semfeat. When dealing with a large number of visual concepts, as presented in Chapter 3, there will be an inherent number of redundant concepts. While we do not have a guarantee that reducing redundancy in a large concept collection will lead to better performing *Semfeat* descriptors, it is a direction worth pursuing. For this goal, we can either pursue a data driven clustering approach [23] or exploit prior knowledge of a semantic hierarchy [104].

Semfeat with concept localization. In the *Semfeat* extraction framework presented in Chapter 4 we give the whole image as input for the visual concept classifiers. While it will undoubtedly add complexity and increase the Semfeat extraction time, introducing an object detection process, such as presented in [381] may lead to better semantic features.

Semfeat for image classification. In this Thesis, we investigated the use of semantic features for content based image retrieval (CBIR). As a complement to CBIR, it is natural to evaluate *Semfeat* performances in an image classification task. In the final Section of Chapter 4, we presented a first experiment towards this direction on the publicly available Pascal VOC 2007 dataset [353]. Clearly, the preliminary results, an Overfeat based version of *Semfeat* offered comparable results to the original *Overfeat* descriptor, but further investigation is actually required. While one of the main advantages of *Semfeat* reside in the capacity to sparsify the descriptor and the use of an inverted index for fast retrieval, adapting *Semfeat* for classification purposes is certainly a direction worth pursuing.

Domain specific credibility Although the dataset that we introduced in Chapter 5 is also designed to allow a fine-grained topic specific credibility analysis of Flickr users, this very task is left for future work. When doing retrieval, one possible way of taking into consideration the topical expertise of a user is by deriving his or her visual credibility estimator by guarding only the predictions from the binary visual classifiers that are semantically close to the query. Besides the credibility features presented in this paper, other credibility descriptors may be extracted from the image metadata but also from other data sources (e.g. user contacts, image comments, groups). Also, mainly due to space constraints, an in-depth analysis of feature importance and feature selection for both proposed tasks could be a promising direction for future work.

Improved credibility estimates for diversity. In Chapter 5 we proposed 66 estimates for user tagging credibility, alongside with methods for feature selection. In the first part of Chapter 6, we tested only the visual credibility estimator in the proposed image retrieval result diversification framework. The next step will certainly be to use the insight we gained from the experiments performed in Chapter 5 and apply learned credibility estimates to the image diversification task.

Credibility estimates in different retrieval scenarios. In Chapter 6, we proposed a framework for adding user credibility estimates to a image retrieval result diversification pipeline. While this has been proven the be efficient in terms of diversification, the use of credibility in more sophisticated retrieval schemes ([32], [33]) is undoubtedly worth investigating.

Putting it all together. In Chapter 3, we evaluated three methods for re-ranking Flickr group images in order to reduce the level of noise. Then, in Chapter 5, we used the visual concept models trained on a subset of these images to create a user tagging visual credibility estimator.

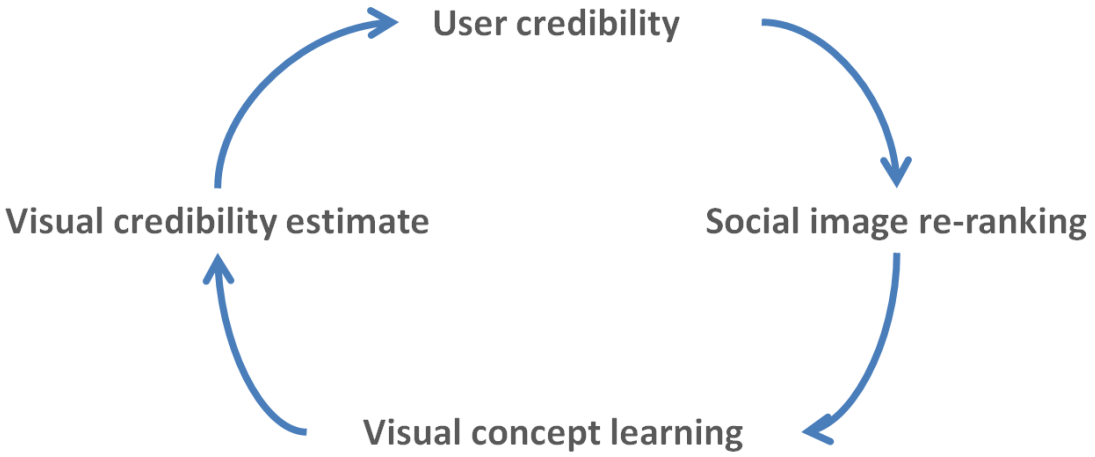


FIGURE 7.1: User credibility and visual concept learning improvement cycle.

As illustrated in Figure 7.1, an obvious alternative for image ranking is to use the credibility of the users who uploaded the images found in Flickr groups. In theory, this should lead to better visual concept models. Moreover, these models can be exploited to produce improved visual credibility estimates. It is immediately noticeable that we are faced with a cyclic improvement process. In some future work, at least one iteration is clearly worth to try.

Appendix A

User Profiling for Answer Quality Assessment in Q&A Communities

A.1 Introduction

The large data volumes shared on collaborative Web applications represent a rich and valuable source of knowledge for the users' daily activities. As proved by the huge success of search engines, users need to rapidly find their way through the plethora of available information. Trust is one key concept operationalized by information retrieval algorithms such as PageRank [255] or HITS [382], which combine statistical matching between user queries and Web page content and the centrality the Web pages in the Web graph. Closely related to trust is the problem of automatically discovering experts, i.e. contributors that are likely to provide valuable contributions to a community of interest. A successful expert detection enables the ranking of contributions based on their quality so as to put forward those that are most likely to be useful. Consequently, the access to relevant pieces of information from the large volume available on the Web is accelerated and the users have their information needs satisfied quicker.

Fast access to relevant information is particularly important when complex information needs, such as learning about a new topic or solving a specific problem, are expressed. When such needs occur, people often consult relevant Web communities (forums, Q&A websites etc.) which gather contributions from a large array of users with different levels of expertise. We place ourselves in a scenario in which the user's information need was met and helpful answers are already available but scattered in large volumes of information of variable quality. Past research on Community Question Answering (CQA) has focused on key aspects such as: expert identification [284], importance of temporal cues [285], [383], combining graph-based and user features [283] or human

factors that contribute to the success of CQA [384] but numerous questions still need to receive appropriate answers. In this work, we focus on the automatic assessment of CQA activity and address the following research questions:

- does coarse grained user profile information contain useful hints for expertise discovery?
- can topic discovery in past contributions be used to predict the quality of new answers?
- can profile and activity data be effectively combined in an automatic answer reranking scenario?

In this work, we present an approach that answers the aforementioned questions. A large dump provided by Stackoverflow, an active forum focused on technical questions, is used for experiments. An analysis of user profiles highlights interesting features that contribute to discriminating expert users is first presented. Then we discuss two application scenarios: quality prediction for newly arrived answers and an automatic answer reranking. Thorough evaluations using Stackoverflow content are proposed for both scenarios. The evaluation results show that the methods introduced in the work significantly outperform appropriate baselines.

A.2 Related Work

Our work builds on the recent research in CQA, which exploits both community structure and user features to automatically detect experts and spot high-quality answers and questions. As we have mentioned, trust is an essential component of successful IR algorithms such as PageRank [255] or HITS [382] and variants of these algorithms focused on domain and user expertise have been proposed. For instance, [385] added domain specificity to link analysis, while [386] applied such algorithms to expert detection in online forums.

[282] argue that an empirical distinction between expert and non-expert contributors to a CQA hampers the overall quality of expert detection. Using a graph based view of the community, they introduce a principled model for authority scores that is based on a mixture of gamma distributions. Then they show that this model is well fitted for the problem posed. Besides their focus on the expert detection, one important difference with our work is that [282] only analyze graph properties to determine expertise whereas we look at profile information, activity data and question metadata. [283] report that

adding domain expertise and user reputation to graph-based features improves expert identification in CQA. However, they only exploit votes given to a user's answers to derive domain expertise and reputation and disregard other relevant user data such as demographic factors or completeness of self-description. [284] discuss the shortcomings of supervised expert detection approaches (i.e. availability of a large set of labeled data) and introduce a semi-supervised method based on coupled mutual reinforcement. Their framework is capable of finding high-quality answers, questions as well as experts by combining a comprehensive array of question, answer and user features. The main differences with our work come from the methods chosen to model users, the use of question metadata and a part of user features (temporal, demographic, self-descriptions) that are not explicitly taken into account by [284].

[285] analyze answerer behavior to determine when and how answers are generated. They confirm that users have daily and weekly periodicities but also point out that there are bursty patterns of activity. Equally interesting, users have favorite categories in which they provide answers but the choice of the questions they answer is mostly determined by their rank in the list of available questions. In a related study, [286] show that expert and non-expert CQA contributors can be differentiated based on a selection bias that is stable over time. Experts tend to choose questions for which they have a chance to make a valuable contribution.

Stackoverflow is a successful community and is increasingly used to support CQA-related research due to the immediate availability of its content but - probably - also to the familiarity of computer scientists with it. [387] tried to match problem difficulty and expertise in order to obtain an efficient distribution of a community's resources. The time lapse between the question and an accepted answer was used as a proxy for query difficulty. Expertise was calculated with simple measures, such as the percentage of positive/negative votes on a user's answer, and a strong correlation was found with the expertise level provided by Stackoverflow. While interesting, this research is incipient and the modeling of problem difficulty needs to be refined. [384] argue that the sustained involvement of the Stackoverflow design team in the community, combined with a good quality technical design, explains the success of this CQA. More closely related to our work [388] and [383] analyze temporal cues that contribute to expert identification and discuss their influence in community dynamics. One important finding reported is that their temporal cues based method outperforms user statistics based ones. With the proposed algorithm, experts can be identified after only 20 weeks of activity in the community. The proposed approach is very interesting but it focuses only on expert identification and not on answer quality prediction and answer ranking.

A.3 User Analysis

Our first goal is to analyze different user profile dimensions in order to observe correlations between them and expertise. We focus on those dimensions that provide a useful separation between expert and non-expert contributors.

A.3.1 Dataset

Stackoverflow is one of the most active and popular CQA that covers a wide area of computer science topics. For the experiments described in this work, we used the August 31 2012 dataset. It contains a total of 2,012,348 questions, 4,456,287 answers and 279,817 unique answerers. We select users that provided at least 10 answers before July 1 2012. These restrictions are set up to keep only users with sufficient contributions and to have a two months frame in which the answers could receive votes. This filtering resulted in a set of 75,657 users. The users were then ranked in ascending order by the average score they received for their answers. These average scores are an expression of the quality of a contributor's activity in the community. The overall average is 1.576, with a standard deviation of 1.84 and minimum and maximum values of -0.88 and 161.5. For the analysis, the data was partitioned in 10% intervals. Other partitions were tried but conclusions similar to those reported here were found.

A.3.2 User Profile Information

A.3.2.1 User name.

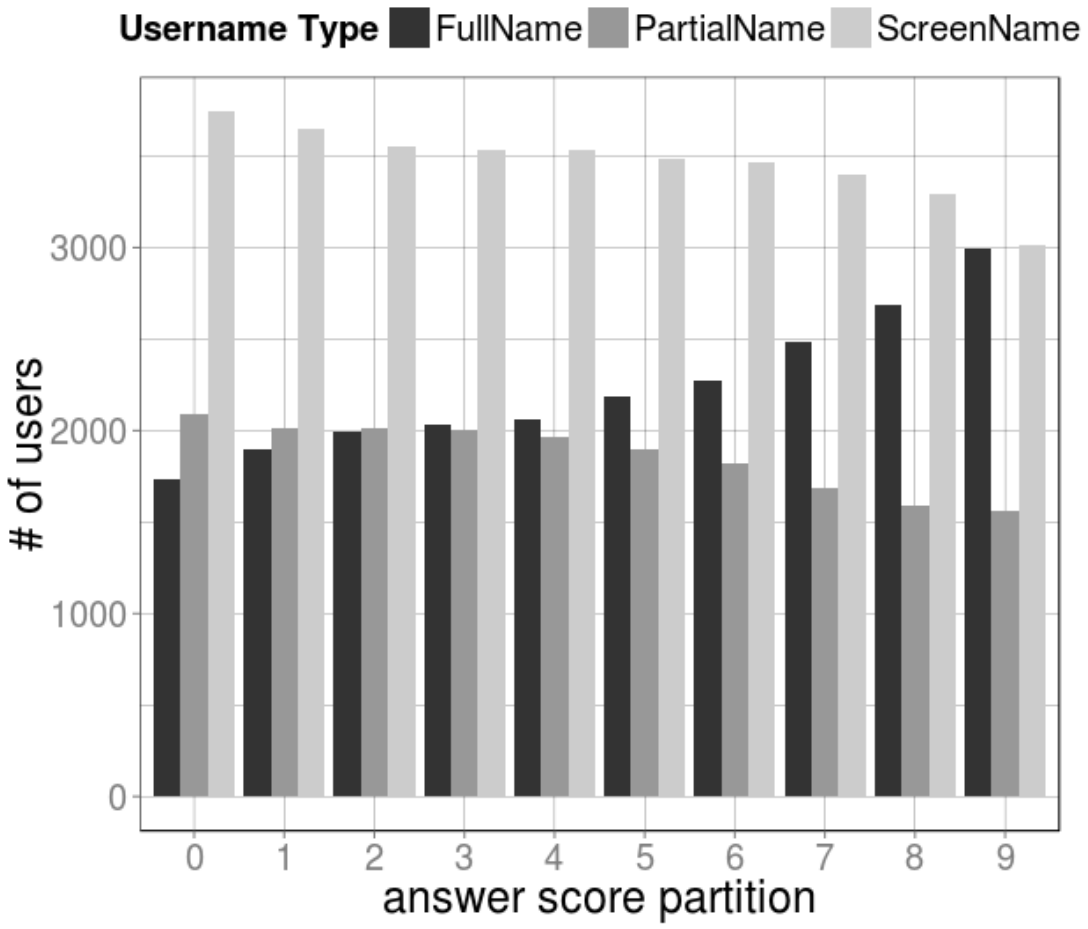


FIGURE A.1: User name types distribution. 0 and 9 stand for the 10% lowest and highest scoring partitions.

According to [389], user name characteristics, such as gender and type (individual or company), indicate the existence of a bias among microblogging users when evaluating content generated by others. Here, we perform a user name analysis in terms of the resemblance to a legal name. We distinguish three types of user names:

- *fullName*: A user name that has a structure resemblant to a legal name. It is composed of at least two words that start with a capital letter, with the exception of particles (e.g. *von, de*) and does not contain numbers.
- *singleName*: A single word that starts with a capital letter and does not contain numbers.

- *screenName*: A single word or multiple words that do not start with a capital letter or contain numbers or other characters.

The results in Figure A.1 indicate that fullName is well correlated with high quality answerers and that, inversely, screenName and singleName are often associated with low quality contributions. These findings show that a cue as simple as the user name is potentially useful to differentiate between expert and non-expert contributors.

A.3.2.2 Self Description.

A total of 40174 users have provided self descriptions which were modeled using a standard tf-idf model. From the 10% user partitions, we select the partitions with the lowest and the highest average answer scores and term category scores were obtained by summing up individual tf-idf scores. To determine specific terms for each group, we select the top 100 terms obtained with tf-idf and then calculate the probability of appearance in each of the two partitions. In Figure A.2, we present the top 20 terms that are most specific to each of the two categories. Terms that are prevalent in the low score partition often pertain to Web related technologies (*php*, *ajax*, *jquery* or *css*) and to databases (*mysql*, *oracle*). Very well represented in this category is *India*, a term whose presence could be explained by the recent development of computer science in this country and by the presence of a large number of Indian students on Stackoverflow (the terms *learn*, *enthusiast* are also prominent in the same partition). High scoring contributors often use terms that indicate that they are established computer science specialists (*engineer*, *developer* or *programmer*) and also declare competencies in programming languages (*python* or *java*).

FIGURE A.2: Most frequent terms for 10% of users with lowest and highest answer scores (left, respectively right).

A.3.2.3 Age.

55.24% of selected users indicated their age. The average age of the 10% lowest scoring partition is 28.14 and that of the highest scoring one 32.6. Intermediate expertise partitions range monotonically between these two values. A similar relationship can be observed when looking at the total number of users from each score interval that have indicated their age. Users that are more comfortable in revealing their age tend to provide more valuable answers.

A.3.2.4 Links to external platforms.

Stackoverflow contributors often provide links towards external Web platforms, such as personal blogs, Twitter or Facebook accounts. In Table A.1 we present statistics that associate the 10 most frequently used platforms with the average score obtained by the users of these platforms. Interestingly, the extreme values are obtained for Twitter (1.89) and Facebook (1.09), two of the most popular social networks. A possible explanation of this difference is that the computer science community might be better represented on Twitter than on Facebook. As expected, high average scores are obtained by users that point toward specialized communities (such as Stackoverflow itself or Github). Scores that are close to the average (1.576) were obtained for most other platforms. A limitation here is that we only look at the links and not at the content of the linked pages. Although potentially relevant, personal websites that are not part of a larger platform cannot thus be analyzed and we leave content analysis for future work.

TABLE A.1: Most frequent platforms. The overall average score is 1.576.

Website	Occurrences	Average Answer Score
Blogspot	2559	1.59
Wordpress	1254	1.53
Stackoverflow	1010	1.8
Twitter	**891**	**1.89**
Google	856	1.53
Linkedin	744	1.62
Github	562	1.86
Facebook	**259**	**1.09**
About	216	1.75
Tumblr	168	1.56

A.3.2.5 Avatars.

If a Stackoverflow user does not upload her own avatar a default one is based on an MD5 hash of the user's email address. The differences between default and personalized avatars are often given by the clarity of the default colors and the variety of shades. A classical image descriptor ($4 \cdot 4 \cdot 10$ multidimensional HSV histograms) that captures such features is combined with a Jeffrey distance to compute image similarities [390]. We manually labeled 2000 images for each class to classify avatars as personalized or default and we obtained an accuracy of 97.45% in a two-fold cross-validation. We then used the complete set of 4000 labeled images as training for the labeling of the remaining avatars. The results presented in Table A.2 show that an average score of 1.743 is obtained for users that personalized their avatar, against 1.373 for the others. This difference indicates that the presence of a personalized avatar is often associated with good quality answers. A distribution of the users that have personalized avatars in the 10% groups is presented in Figure A.3.

A.3.2.6 Other features.

We have tested a number of other features and present results for location and existence of a link in the profile in Figure A.3. A monotonic increase of the number of users that have informed these two dimensions of their profile with the average answer quality score is observed.

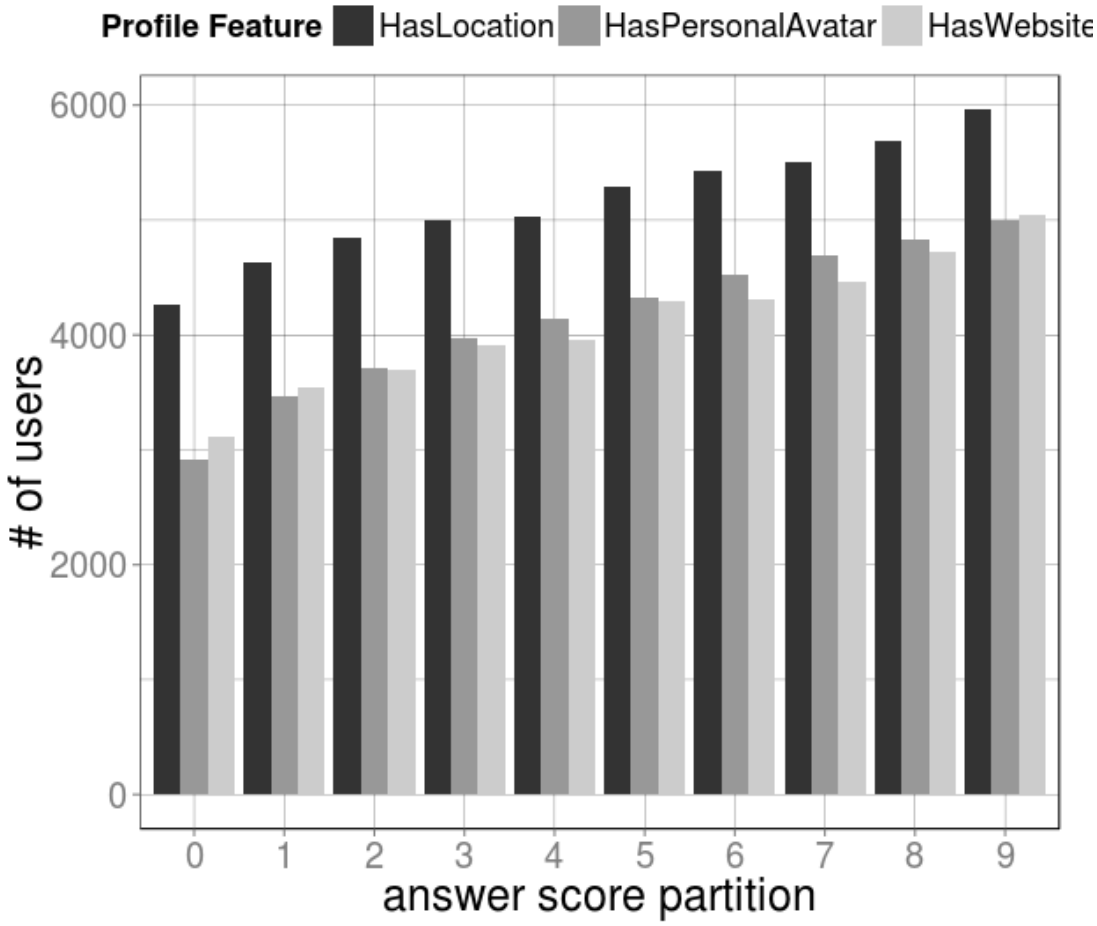


FIGURE A.3: Distribution of supplementary profile features.

A.3.2.7 Overview.

TABLE A.2: Overview of user profile dimensions.

Profile Feature	Value	# of users	Score
Location	Present	51654	1.687
	Absent	24002	1.338
Description	Present	40174	1.676
	Absent	35482	1.45
User name	FullName	22345	1.808
	PartialName	18639	1.46
	ScreenName	34672	1.49
Avatar	Default	34057	1.373
	Personal	41599	1.743
Website	Absent	34596	1.383
	MalformedURL	1700	1.458
	Down	4101	1.627
	Active	35259	1.766

Table A.2 summarizes our findings concerning some of the most interesting user profile dimensions. Globally, the more complete a given user profile is, the higher the probability to obtain good quality answers from that user is. Of particular interest are features such as location, user name and avatar, whose presence results in scores that are sensibly higher than the average user score (1.576).

A.3.3 User Community Involvement

We examine the involvement of Stackoverflow users in the community and present statistics about positive and negative votes they provide, as well as about the number of profile views in Figure A.4. The obtained results clearly show that community involvement is correlated with expertise for the three analyzed user activity cues. This is particularly true for positive votes and for the number of views, for which users with important activity are strongly concentrated in the best four 10% groups (6 to 9 in Figure A.4).

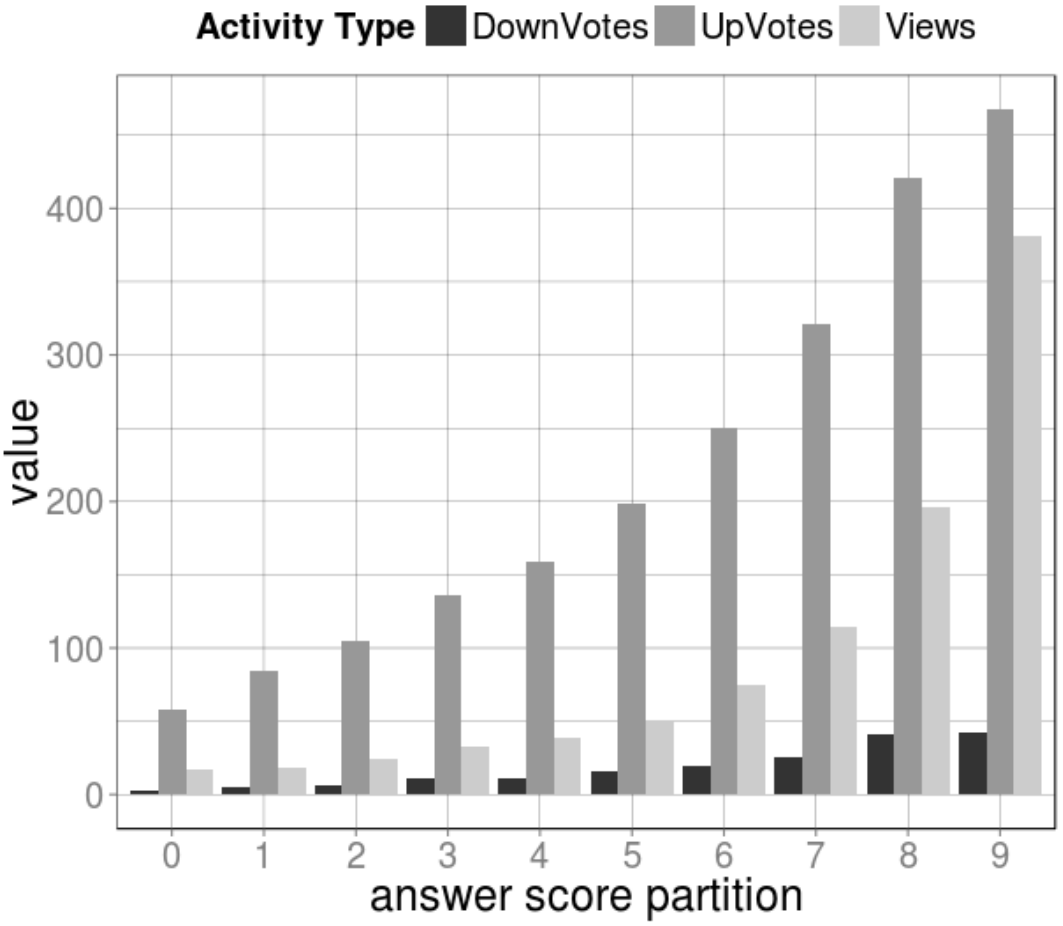


FIGURE A.4: Community involvement features over the answer score partitions. 0 and 9 stand for the 10% lowest and highest scoring partitions.

A.4 Answer Quality Prediction

Automatic answer quality prediction is important in order to provide new answer ranking if not enough community votes are available. We introduce a supervised approach to quality prediction that takes into account past user activity. The topics of interest of a user that arise from the questions she answers are modeled using Latent Dirichlet Allocation (LDA) [391] and Explicit Semantic Analysis (ESA) [392]. These two models are complementary because LDA models text properties based a small number of hidden topics, while ESA provides an explicit mapping of examined documents onto a predefined concept space. In addition to LDA and ESA, simple bag of words models were tried but they yielded poor results that are not discussed hereafter. Each question asked has at least one associated tag that can range from a specific product to a broad computer science domain. These tags are used as an indirect representation of a user's answers.

A.4.1 LDA-based Topic Modeling

We use LDA [391] to assign topics to questions. The interpretable topic distributions appear by computing a hidden structure generated from the observed collection of instances [393]. Topic models are appropriate because that they do not require a priori knowledge of data, such a tag taxonomy. Due to the changing nature of collaborative tagging, this property is a key aspect for the practicality of topic detection in our scenario. Also, the method is robust to spelling mistakes, common in social tagging. For our experiments, we used the Mallet LDA implementation [394]. The framework uses Gibbs sampling for constructing the sample distributions that are exploited for the creation of the topic models. The models are built using the list of tags associated with a question as instances. In the answer quality prediction experiments, each LDA instance is encoded by inferring a topic distribution. Given a set of tags of any size, if a model with 20 topics is used, we generate a vector of length 20, where each the i-th value is the probability that the list of tags belongs to the i-th topic.

A.4.2 ESA-based Topic Modeling

ESA was introduced in order to exploit the collective intelligence of Wikipedia editors in tasks such as word relatedness or document classification [392]. A set of support concepts (Wikipedia articles) is modeled using tf-idf and an inversed index that maps words onto these concepts is produced. Given an entry text, concept representations of individual words in that text are summed up to produce a aggregated representation of the text. Then, the similarity between two texts can be calculated using similarity measures such as the cosine similarity. There are several available implementations of ESA but, since the fine tuning details of the method were only recently published, their performances are reduced compared to the original implementation [392]. Following the publication of method parameters, we have implemented an ESA version whose performances on the WordSim-353 dataset are close to those given by [392] (0.73 vs 0.75 Spearman rank-order correlation using the November 2005 Wikipedia dump). Here we compute ESA vectors based on the September 2012 English Wikipedia dump. The tags associated to questions are projected onto the ESA concept space and this mapping is used to predict the quality of new answers.

A.4.3 Experiments

We randomly select 100 users that provided at least 510 answers before July 2012. The time constraint has the same role as in user analysis, namely allowing a two months

period in which the answers could receive votes. The answers are listed in a chronological order. The most recent 10 answers of each selected contributor are used for test. We use subsets of N answers ranging between 10 to 500 items with a step of 10, to derive different training partitions. The most recent N answers in each partition that precede the 10 test answers are kept. As a baseline, we propose to associate the average score of past answers to any new answer. This measure assumes that we do not have any information about the topics of previous answers.

Experiments with both LDA and ESA were carried using a weighted cosine similarity prediction, detailed in equation A.1 and a SVM classifier with a Radial Basis kernel. The loss function of the classifier is well adapted for our answer quality prediction which was modeled as a bound-constraint regression problem. During LDA tuning experiments, we also tested a Random Forests (RF) classifier. Its poor performances determined us to drop it in the final experiments.

$$WCos = \frac{\sum_k cos(\overrightarrow{a_i}, \overrightarrow{a_k}) \cdot score(a_k)}{\sum_k cos(\overrightarrow{a_i}, \overrightarrow{a_k})} \tag{A.1}$$

, where $\overrightarrow{a_i}$ is the vectorial representation of a test instance, $\overrightarrow{a_k}$ is the vectorial representation of a training instance and $score(a_k)$ is the community score for the answer. k is the number of previous answers that are taken into account. We perform tests with a value of k ranging from 10 to 500 with an incremental step of 10.

To test the proposed methods, we first compute the Root Mean Squared Error (RMSE) between predicted and actual values of the test instances. We then evaluate the methods by averaging RMSE over the entire user test dataset.

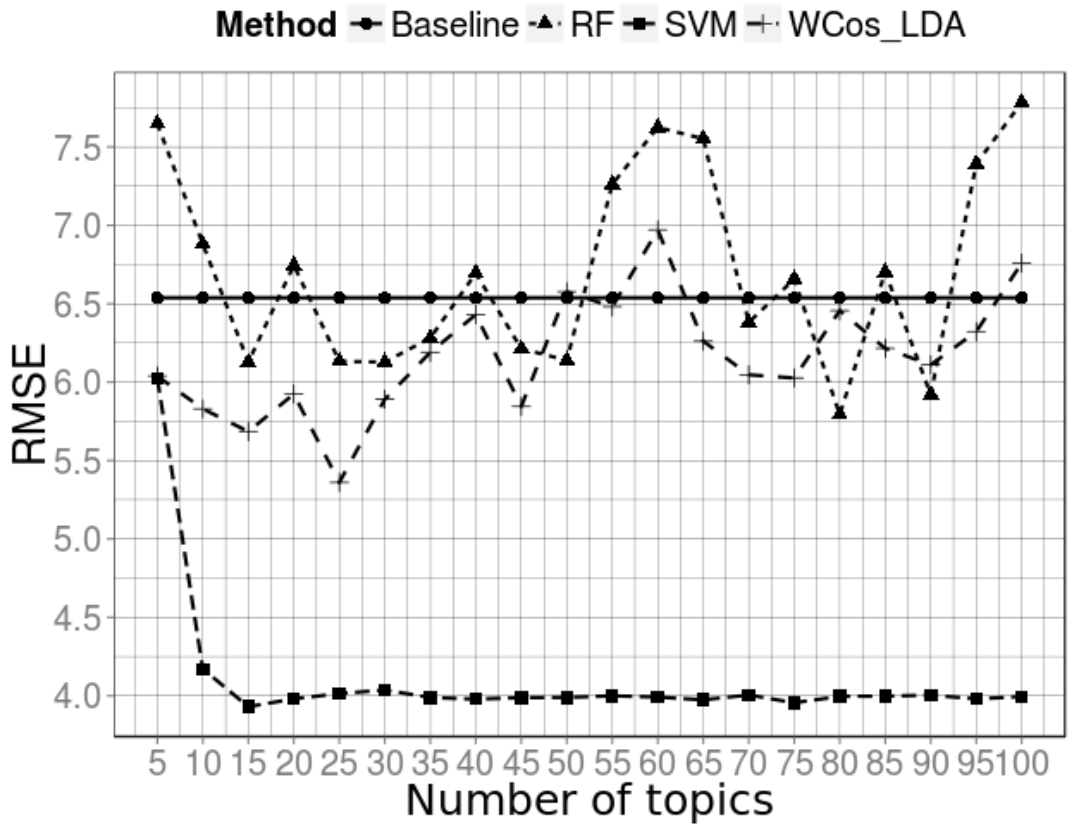


FIGURE A.5: Topic number influence.

The most important parameter of LDA is the number of model topics. In Figure A.5, we compare the variation of performances for three different methods with a number of topics between 5 and 100. The number of training instances was fixed at 100 for each user. WCos and RF methods have poor performances, comparable to those of the baseline. The SVM based method is consistently more accurate and the number of topics has a strong influence on performances for values up to 20 but tends to yield stable results past this value. Consequently, the other experiments that involve LDA modeling use 20 topics.

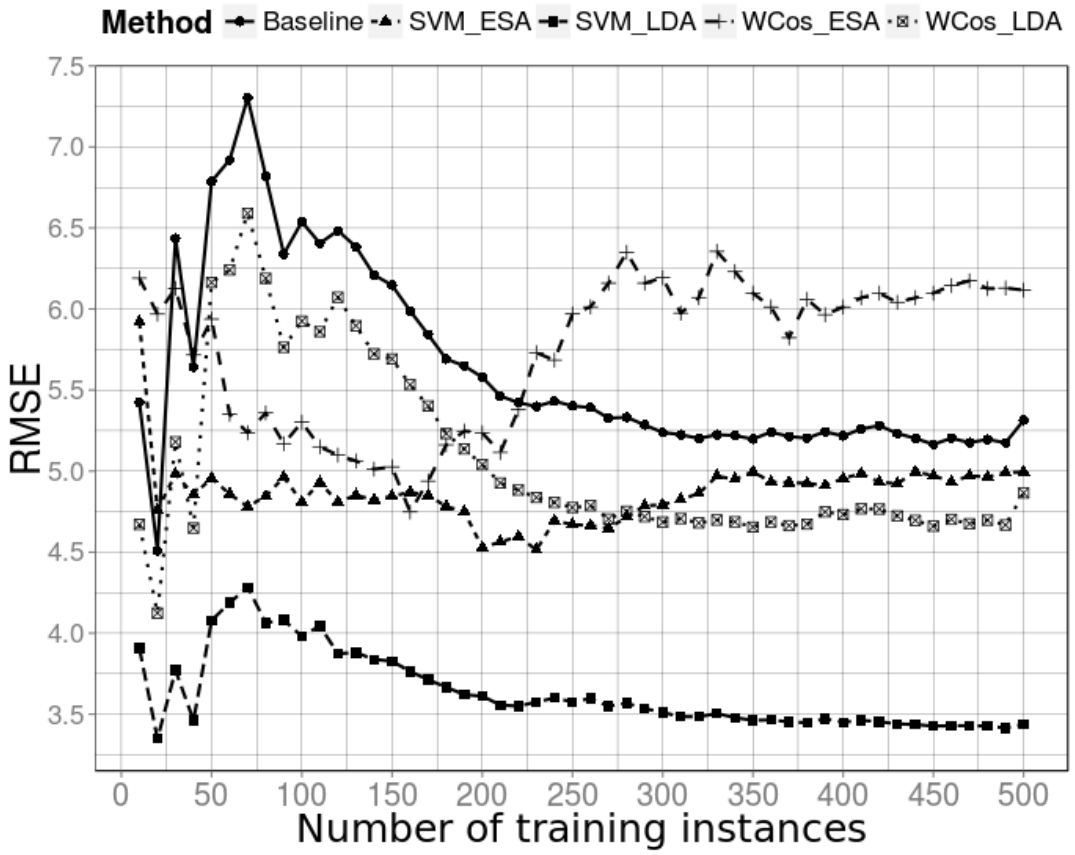


FIGURE A.6: Influence of the number of training instances. The smaller the RMSE value is, the better the performances are.

The results presented in Figure A.6 show that the LDA based methods outperform the ESA based ones and the baseline. The comparison of weighted cosine and SVM based methods for score prediction confirms the intuition that the latter more sophisticated method is more accurate. As expected, the variation of the number of training instances has an important impact up to 300 instances. Except for WCos_ESA, a plateau is reached after this value signifying that enough training data are available. The best results are obtained with SVM_LDA, a method that is consistently better than all the others regardless of the number of training instances. For instance, at 450 instances, SVM_LDA reduces the RMSE by approximately 33% compared to the baseline (3.45 vs. 5.2) and by 30% compared to SVM_ESA (3.45 vs. 4.97). The good SVM_LDA performances obtained for a small number of training instances show important predictive power even confronted to data scarcity. Surprisingly, the best results with ESA based method are obtained when around 200 instances are available for training. This behavior is not fully explained and needs further checking. Interestingly, the baseline produces results that are in the same range as SVM_ESA and WCos_LDA. This result advocates

for an appropriate choice of both the document modeling method and prediction model to use.

The results presented here demonstrate the suitability of combining LDA for modeling question metadata and SVM for score prediction in the proposed setting. LDA topic models were obtained from domain training data and are thus well correlated with the test data. The results obtained with ESA are somewhat deceiving but they can be explained by the fact that no domain adaptation of the method was performed. Equally important, a significant number of tags associated to the questions used (1384 out of 5017) could not be matched to the ESA vocabulary. Although basic tag preprocessing (hyphen or underscore elimination, stemming) was performed, more advanced methods need to be devised. Tackling these two problems would probably improve the performances of ESA based methods but falls outside of the immediate scope of this work. For both LDA and ESA, only coarse grained metadata (i.e. tags associated to the questions) are used in order to test their effectiveness when only indirect answer information is available. In future work, we will add the content of the answers to the models in order to have more accurate representations of the data that is analyzed and to improve the overall performance of the prediction process.

A.5 Automatic Answer Ranking

In CQAs there are no hired experts to provide answers or control them. Answer quality is derived from the feedback offered by community members and it is usually validated over time. Although a question may be quickly answered, the number of votes its answers receive is influenced by the popularity of the topic or the quality of the question [384].

Answer ranking is essential for providing fast access to the best available content. When the community is active enough, as it is the case with Stackoverflow, good quality rankings are obtained based on the users' votes. In other CQAs, insufficient feedback is available and efficient automatic ranking schemes are needed. We present a series of methods that leverage user profile information, activity data or a combination of the two in order to obtain automatic answer rankings. In addition, a method based on answer quality prediction is presented.

A.5.1 Ranking Methods

Stackoverflow questions receive an average of 2.1 answers. Automatic ranking methods, whose role is to facilitate access to relevant pieces of knowledge, are particularly useful

when a larger number of answers is available. To account for this, we randomly picked 100,000 questions that have between 5 and 10 answers. The lower bound allows a clearer separation between the different methods tested. The upper bound is used to create a uniform test dataset.

Except for the answer quality prediction method, we use different combinations of features derived from user profile information or activity to train a Ranking Support Vector Machine (RSVM) classifier [395] that is specially designed for ranking problems. It learns a linear classification rule that optimizes the loss function represented by the total number of swapped pairs in the rankings. Continuous features were normalized and categorical features, such as the user name, were encoded as vectors. For example, a *fullName* type is represented as $< 1, 0, 0 >$. We also test a ranking by the answer score predicted by SVM_LDA, the method that provided the best results in the answer scoring experiments. Next, we detail the test configurations:

- *RSVM_User_Profile* - this configuration exploits all user profile dimensions presented in Table A.2.
- *RSVM_User_Profile_Selected* - different user feature combinations were tested based on the observations made during profile analysis. The results reported in Table A.3 correspond to best combination of features, namely *user name*, *avatar* and *location*.
- *RSVM_User_Activity* - this configuration exploits the number of up/down votes provided by the user and the profile views presented in Figure A.4.
- *RSVM_User_Mixed* - a combination of features used by *RSVM_User_Profile_Selected* and *RSVM_User_Activity*.
- *RSVM_User_Descriptions* - this configuration exploits user descriptions which are modeled as Bag of Words with tf-idf weighting applied to lemmatized versions of the words. Given their high variability and potential usefulness for detecting reliable users, all the URLs are replaced by the word *link* in order to be able to match them.
- *Topic_Answer_Score_Prediction* - each answer is ranked based on its predicted quality. The SVM_LDA with 20 topics is used here for answer ranking. Considering the plateau reached by the SVM_LDA classifier after 300 training instances (Figure A.6), we retain a maximum of 300 most recent answers for training. To avoid the cold start problem (i.e. the user's first answer), we use the global average answer score.

A.5.2 Experiments

We partition the data (100,000 answer sets) in 80/10/10 splits. We use the first 80% for training and tunning the parameters of the RSVM classifier, the following 10% for validation and we report the results on the final 10%.

We compare the different answer ranking methods presented above with two baselines:

- *Random_ Presentation* - results are presented randomly. The results in Table A.3 are obtained by averaging 10 random rankings.
- *Temporal_ Order* - results are presented using the order in which they arrive on Stackoverflow.

Two complementary metrics were used during experiments:

- *Levenshtein distance* [396] (LEV in Table A.3) between an ideal ranking, in which answers are sorted by descending Stackoverflow score, and the automatic rankings. Roughly speaking, this distance gives the minimum number of items to change needed in order to transform a string into another and is a good proxy for string similarity. To calculate it, answer scores are cast as items and the rankings as strings. In Table A.3, the smaller the Levenshtein distance is, the better the results are.
- *Best answer position* (BAP in Table A.3). Indicates the number of answers a user has to go through before finding the best one.

The first metric gives an overview of the global quality of answer ranking by comparing them to an ideal one. The second metric is local and estimates the number of answers a user has to go through before finding the best one.

TABLE A.3: Comparison of different answer ranking methods.

Method	LEV	BAP
Random_Presentation	4.03	3.12
Temporal_Order	3.69	2.61
RSVM_User_Profile	3.56	3.29
RSVM_User_Profile_Selected	3.15	2.84
RSVM_User_Activity	3.23	2.66
RSVM_User_Mixed	**2.87**	**2.43**
RSVM_User_Descriptions	3.08	2.54
Topic_Answer_Score_Prediction	3.38	2.69

As expected, the Temporal_Order baseline is better than a Random_Presentation and will be used for comparison with automatic configurations. The results in Table A.3 show that all automatic answer rankings outperform Temporal_Order for a global assessment with the Levenshtein distance. Particularly interesting results are obtained for RSVM_User_Mixed, a configuration that combines selected profile features and user activity. An average 22% reduction of the Levenshtein distance from 3.69 vs. 2.87 is obtained. This performance is probably a result of the complementarity of the profile and activity data. Good global performances are also obtained with RSVM_User_Profile_Selected (3.15) and RSVM_User_Descriptions (3.08).

Results are more mixed for best answer position. Slight improvements are obtained with RSVM_User_Mixed (2.43) and RSVM_User_Descriptions (2.54) compared to Temporal_Order (2.61). The very active Stackoverflow community makes the temporal order hard to beat since best answers to simple questions are obtained very quickly.

Encouraging results are equally obtained for the ranking based on answer quality score prediction - 8.4% improvement of the global ranking with respect to the baseline. However, Topic_Answer_Score_Prediction lags behind most of user profile and activity based methods.

The results presented here show that good performances can be achieved by combining user activity and profile information. Interestingly, short user descriptions (RSVM_User_Descriptions) can also be leveraged to rank answers. A fusion of RSVM_User_Mixed, RSVM_User_Descriptions and Topic_Answer_Score_Prediction is likely to further improve the quality of the results. However, given the practical difficulties posed by the representation of the features used by these methods in a common space, it is beyond the scope of the present work. These results are in line with past findings reported in [284] or [383] using different features and learning methods. Except for Topic_Answer_Score_Prediction, the proposed methods do not depend on interaction data, such as answer scores, and open the way for automatic answer ranking on collaborative platforms that do not receive as much user feedback as Stackoverflow.

A.6 Conclusions and Future Work

Forums and CQAs are privileged venues for sharing and retrieving useful information. Their usefulness is well established but, with the large amounts of information available, automatic processing methods are needed to accelerate the users' access to relevant content. Related to our initial research questions, this work contributes to the following hot topics in CQA:

- Improved understanding of CQA content through a detailed user profile analysis. Interesting correlations between answer quality and the tested features were found. As expected, the more complete a profile is, the higher the chances are for that user to provide good quality answers. However, not all features are equally discriminant and we have shown that user name, user location and personalized avatars are the most useful ones in an answer ranking task.

- Innovative application of LDA and ESA, two well-established text representation methods, provides good results in an answer quality prediction scenario. Compared to a baseline that attributes quality score based on the average of existing answer, the answer quality prediction is improved by 33% by the best proposed method.

- Proposition of novel automatic answer ranking methods. Different configurations, based on user profile information and/or user activity were proposed. We showed that an appropriate combination of the two types of data gives the most accurate results. A 22% performance boost is obtained compared to a ranking based on the temporal order of the answers. Encouraging results are equally obtained for the use of answer scoring in the answer ranking scheme.

These contributions advance the state of the art in CQA related studies from a methodological point of view, with the innovative use of machine learning and text representation techniques. From a data leveraging perspective, we innovate through the use of profile dimensions that were ignored or marginally considered in previous works (i.e. user avatar, user name, location etc.)

While here we focused on coarse grained data that are easily tractable, in future work we will concentrate on the introduction of more detailed features, extracted from the training answers but also from external resources (i.e. Webpages that are linked in the user profiles). Their introduction is likely to significantly improve the quality of answer scoring and ranking. We will first work towards the extension of answer scoring models with answer texts. We will also create a domain related version of ESA and devise more advanced methods for reducing the vocabulary mismatch between Stackoverflow and Wikipedia. Then, we will explore late fusion approaches for combining different answer ranking methods presented in this work. Finally, while here we proposed ranking methods that assume answer independence, it is important to study the dynamics and interactions of contributions and we will incorporate them in the proposed framework.

Appendix B

Publications

2015

[1] **Alexandru Lucian Ginsca**, Adrian Popescu, and Mihai Lupu. Credibility in Information Retrieval. *Foundations and Trends in Information Retrieval 9.5 (2015)*, pages 1-107.

[2] Bogdan Ionescu, **Alexandru Lucian Ginsca**, Bogdan Boteanu, Adrian Popescu, Mihai Lupu, and Henning Müller. Retrieving Diverse Social Images at MediaEval 2015: Challenge, Dataset and Evaluation. In *MediaEval 2015 Workshop*, Wurzen, Germany, 2015.

[3] **Alexandru Lucian Ginsca**, Adrian Popescu, Mihai Lupu, Adrian Iftene, and Ioannis Kanellos. Evaluating User Image Tagging Credibility. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2015)*, pages 41-52. Springer International Publishing, 2015.

[4] **Alexandru Lucian Ginsca**, Adrian Popescu, Hervé Le Borgne, Nicolas Ballas, Phong Vo, and Ioannis Kanellos. Large-Scale Image Mining with Flickr Groups. In *MultiMedia Modeling*. Springer, pages 318–334, 2015. **(Best paper)**

[5] Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, **Alexandru Lucian Ginsca**, Adrian Popescu, Yiannis Kompatsiaris, and Ioannis Vlahavas. Improving Diversity in Image Search via Supervised Relevance Scoring. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 323-330. ACM, 2015.

[6] Phong D. Vo, **Alexandru Lucian Ginsca**, Hervé Le Borgne, and Adrian Popescu. Effective Training of Convolutional Networks using Noisy Web Images. In *Proceedings of the 13th International Workshop on Content-Based Multimedia Indexing (CBMI)*, IEEE, Prague, The Czech Republic, 2015.

[7] Bogdan Ionescu, Adrian Popescu, Mihai Lupu, **Alexandru Lucian Ginsca**, Bogdan Boteanu, and Henning Müller. Div150Cred: A social image retrieval result diversification with user tagging credibility dataset. *ACM Multimedia Systems-MMSys*, Portland, Oregon, USA, 2015.

2014

[8] **Alexandru Lucian Ginsca**, Adrian Popescu, Bogdan Ionescu, Anil Armagan, and Ioannis Kanellos. Toward an Estimation of User Tagging Credibility for Social Image Retrieval. In *Proceedings of the ACM International Conference on Multimedia (ACMMM 2014)*, pages 1021-1024. ACM, 2014.

[9] Bogdan Ionescu, Adrian Popescu, Mihai Lupu, **Alexandru Lucian Ginsca** and Henning Müller. Retrieving diverse social images at mediaeval 2014: Challenge, dataset and evaluation. In *MediaEval 2014 Workshop*, Barcelona, Spain, 2014.

[10] **Alexandru Lucian Ginsca**, Adrian Popescu, and Navid Rekabsaz. CEA LIST's Participation at the MediaEval 2014 Retrieving Diverse Social Images Task. In *Proceedings of the MediaEval Multimedia Benchmark Workshop*, Barcelona, Spain, 2014.

2013

[11] **Alexandru Lucian Ginsca** and Adrian Popescu. User profiling for answer qualityassessment in Q&A communities". In *Proceedings of the 2013 workshop on Data-driven user behavioral modelling and mining from social media CIKM Workshop*. ACM, pages 25–28, San Francisco, USA, 2013.

[12] Morgane Marchand, **Alexandru Lucian Ginsca**, Romaric Besançon, and Olivier Mesnard. [LVIC-LIMSI]: Using Syntactic Features and Multi-polarity Words for Sentiment Analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 418–424, 2013.

[13] **Alexandru Lucian Ginsca**. Estimating User Credibility in Multimedia Information Flows. In *Fifth BCSIRSG Symposium on Future Directions in Information Access (FDIA 2013)*, Granada, Spain, 2013.

Bibliography

[1] Meng Wang, Bingbing Ni, Xian-Sheng Hua, and Tat-Seng Chua. Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Computing Surveys (CSUR)*, 44(4):25, 2012.

[2] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1349–1380, 2000.

[3] Li jia Li, Hao Su, Li Fei-fei, and Eric P. Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in Neural Information Processing Systems*, pages 1378–1386, 2010.

[4] Alessandro Bergamo, Lorenzo Torresani, and Andrew W Fitzgibbon. Picodes: Learning a compact code for novel-category recognition. In *Advances in Neural Information Processing Systems*, pages 2088–2096, 2011.

[5] Alessandro Bergamo and Lorenzo Torresani. Classemes and other classifier-based features for efficient object categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(10):1988–2001, 2014.

[6] Lyndon S Kennedy, Shih-Fu Chang, and Igor V Kozintsev. To search or to label?: predicting the performance of search-based automatic image classifiers. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 249–258. ACM, 2006.

[7] Hamid Izadinia, Ali Farhadi, Aaron Hertzmann, and Matthew D Hoffman. Image classification and retrieval from user-supplied tags. *arXiv preprint arXiv:1411.6909*, 2014.

[8] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, page 48. ACM, 2009.

[9] Xirong Li, Tiberio Uricchio, Lamberto Ballan, Marco Bertini, Cees GM Snoek, and Alberto Del Bimbo. Socializing the semantic gap: A comparative survey on image tag assignment, refinement and retrieval. *arXiv preprint arXiv:1503.08248*, 2015.

[10] B. Russell and al. Labelme: a database and web-based tool for image annotation. *IJCV*, 77:157–173, 2007.

[11] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International conference on artificial intelligence and statistics*, pages 249–256, 2010.

[12] J. Deng and al. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. of CVPR 2009*.

[13] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, abs/1405.3531, 2014.

[14] Pierre Sermanet and al. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, 2013.

[15] Morgan Ames and Mor Naaman. Why we tag: motivations for annotation in mobile and online media. In *ACM CHI 2007*, pages 971–980.

[16] Pengcheng Wu, Steven Chu-Hong Hoi, Peilin Zhao, and Ying He. Mining social images with distance metric learning for automated image tagging. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 197–206. ACM, 2011.

[17] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *ACM CIVR 2009*.

[18] Tobias Weyand and Bastian Leibe. Visual landmark recognition from internet photo collections: A large-scale evaluation. *Computer Vision and Image Understanding*, 135:1–15, 2015.

[19] Stefan Romberg, Lluis Garcia Pueyo, Rainer Lienhart, and Roelof Van Zwol. Scalable logo recognition in real-world images. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 25. ACM, 2011.

[20] Bogdan Ionescu, Adrian Popescu, Mihai Lupu, Alexandru L Gınsca, and Henning Müller. Retrieving diverse social images at mediaeval 2014: Challenge, dataset and evaluation. In *MediaEval 2014 Workshop, Barcelona, Spain*, 2014.

[21] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015.

[22] Lorenzo Torresani and al. Efficient object category recognition using classemes. In *ECCV*. Springer, 2010.

[23] Alessandro Bergamo and Lorenzo Torresani. Meta-class features for large-scale object categorization on a budget. In *CVPR*. IEEE, 2012.

[24] Stéphane Clinchant and al. Xrce's participation in wikipedia retrieval, medical image modality classification and ad-hoc retrieval tasks of imageclef 2010. In *CLEF'10*.

[25] Hervé Jégou and al. Aggregating local image descriptors into compact codes. *PAMI*, 2012.

[26] Ben Shneiderman. Building trusted social media communities: A research roadmap for promoting credible content. In *Roles, Trust, and Reputation in Social Media Knowledge Markets*, pages 35–43. Springer, 2015.

[27] Julia Schwarz and Meredith Morris. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 1245–1254. ACM, 2011.

[28] Yusuke Yamamoto and Katsumi Tanaka. Enhancing credibility judgment of web search results. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 1235–1244. ACM, 2011.

[29] Alia Amin, Junte Zhang, Henriette Cramer, Lynda Hardman, and Vanessa Evers. The effects of source credibility ratings in a cultural heritage information aggregator. In *Proceedings of the 3rd workshop on Information credibility on the web*, pages 35–42. ACM, 2009.

[30] Nicholas Diakopoulos and Irfan Essa. Modulating video credibility via visualization of quality evaluations. In *Proceedings of the 4th workshop on Information credibility*, pages 75–82. ACM, 2010.

[31] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1717–1724. IEEE, 2014.

[32] N. Jain and al. Experiments in diversifying flickr result sets. In *Proc. of MediaEval Wksp. 2013*.

[33] D. Corney and al. Socialsensor: Finding diverse images at mediaeval 2013. In *Proc. of MediaEval Wksp. 2013*.

[34] Bernardine MC Atkinson. Captology: A critical review. In *Persuasive Technology*, pages 171–182. Springer, 2006.

[35] Carlos Castillo and al. Information credibility on twitter. In *Proc. of WWW 2011*, pages 675–684.

[36] Alexandra Olteanu, Stanislav Peshterliev, Xin Liu, and Karl Aberer. Web credibility: features exploration and credibility prediction. In *Proceedings of the 35th European conference on Advances in Information Retrieval*, ECIR'13, pages 557–568, Berlin, Heidelberg, 2013. Springer-Verlag. ISBN 978-3-642-36972-8. doi: 10.1007/978-3-642-36973-5_47. URL http://dx.doi.org/10.1007/978-3-642-36973-5_47.

[37] Panagiotis G. Ipeirotis and al. Quality management on amazon mechanical turk. In *HCOMP 2010*.

[38] Ling Xu and al. Credibility-oriented ranking of multimedia news based on a material-opinion model. *Web-Age Inf. Mgmt.*, pages 290–301, 2011.

[39] Yusuke Yamamoto and Katsumi Tanaka. Imagealert: credibility analysis of text-image pairs on the web. *SAC 2011*.

[40] B. J. Fogg and Hsiang Tseng. The elements of computer credibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, pages 80–87, New York, NY, USA, 1999. ACM. ISBN 0-201-48559-1. doi: 10.1145/302979.303001. URL http://doi.acm.org/10.1145/302979.303001.

[41] Claudia Hauff, Bart Thomee, and Michele Trevisiol. Working notes for the placing task at mediaeval 2013. In *MediaEval*, 2013.

[42] Devi Parikh and C Lawrence Zitnick. The role of features, algorithms and data in visual recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2328–2335. IEEE, 2010.

[43] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42 (3):145–175, 2001.

[44] Krystian Mikolajczyk and Cordelia Schmid. An affine invariant interest point detector. In *Computer Vision—ECCV 2002*, pages 128–142. Springer, 2002.

[45] Tinne Tuytelaars. Dense interest points. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2281–2288. IEEE, 2010.

[46] Raphael Maree, Pierre Geurts, Justus Piater, and Louis Wehenkel. Random subwindows for robust image classification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 34–40. IEEE, 2005.

[47] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[48] Koen EA Van De Sande, Theo Gevers, and Cees GM Snoek. Evaluating color descriptors for object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1582–1596, 2010.

[49] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer vision–ECCV 2006*, pages 404–417. Springer, 2006.

[50] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[51] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477. IEEE, 2003.

[52] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004.

[53] Josef Sivic and Andrew Zisserman. Efficient visual search of videos cast as text retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31 (4):591–606, 2009.

[54] Jan C Van Gemert, Cor J Veenman, Arnold WM Smeulders, and Jan-Mark Geusebroek. Visual word ambiguity. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(7):1271–1283, 2010.

[55] Y-Lan Boureau, Francis Bach, Yann LeCun, and Jean Ponce. Learning mid-level features for recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2559–2566. IEEE, 2010.

[56] Subhransu Maji and Alexander C Berg. Max-margin additive classifiers for detection. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 40–47. IEEE, 2009.

[57] Gang Wang and al. Learning image similarity from flickr groups using stochastic intersection kernel machines. In *CVPR*, 2009.

[58] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[59] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *Computer Vision–ECCV 2010*, pages 143–156. Springer, 2010.

[60] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *CVPR 2010*. IEEE Computer Society.

[61] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE PAMI*, 34(9):1704–1716, 2012. ISSN 0162-8828. doi: http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.235.

[62] O. Russakovsky and al. Imagenet large scale visual recognition challenge, 2014.

[63] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[64] Florent Perronnin and al. Large-scale image retrieval with compressed fisher vectors. In *CVPR 2010*.

[65] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[66] Quoc V Le. Building high-level features using large scale unsupervised learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8595–8598. IEEE, 2013.

[67] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep fisher networks for large-scale image classification. In *Advances in neural information processing systems*, pages 163–171, 2013.

[68] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014.

[69] Bojan Pepik, Rodrigo Benenson, Tobias Ritschel, and Bernt Schiele. What is holding back convnets for detection? *arXiv preprint arXiv:1508.02844*, 2015.

[70] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jagannath Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*. IEEE.

[71] Xianjie Chen and Alan L Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Advances in Neural Information Processing Systems*, pages 1736–1744, 2014.

[72] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *arXiv preprint arXiv:1503.03832*, 2015.

[73] Hanxi Li, Yi Li, and Fatih Porikli. Robust online visual tracking with a single convolutional neural network. In *Computer Vision–ACCV 2014*, pages 194–209. Springer, 2015.

[74] Philipp Fischer, Alexey Dosovitskiy, and Thomas Brox. Descriptor matching with convolutional neural networks: a comparison to sift. *arXiv preprint arXiv:1405.5769*, 2014.

[75] Jure Žbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. *arXiv preprint arXiv:1409.4326*, 2014.

[76] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. *arXiv preprint arXiv:1504.06852*, 2015.

[77] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. *arXiv preprint arXiv:1504.06375*, 2015.

[78] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *arXiv preprint arXiv:1411.4038*, 2014.

[79] Ali S Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 512–519. IEEE, 2014.

[80] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[81] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[82] Ben Athiwaratkun and Keegan Kang. Feature representation in convolutional neural networks. *arXiv preprint arXiv:1507.02313*, 2015.

[83] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[84] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014.

[85] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *Computer Vision–ECCV 2014*, pages 392–407. Springer, 2014.

[86] James Philbin, Ondřej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[87] Andrej Mikulík, Michal Perdoch, Ondřej Chum, and Jiří Matas. Learning a fine vocabulary. In *Computer Vision–ECCV 2010*, pages 1–14. Springer, 2010.

[88] Jianxiong Xiao, James Hays, Krista Ehinger, Aude Oliva, Antonio Torralba, et al. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010.

[89] Albert Gordoa, Jose A Rodriguez-Serrano, Florent Perronnin, and Ernest Valveny. Leveraging category-level labels for instance-level image retrieval. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3045–3052. IEEE, 2012.

[90] Hervé Jégou and Ondřej Chum. Negative evidences and co-occurences in image retrieval: The benefit of pca and whitening. In *Computer Vision–ECCV 2012*, pages 774–787. Springer, 2012.

[91] Ken Chatfield, Karen Simonyan, and Andrew Zisserman. Efficient on-the-fly category retrieval using convnets and gpus. In *Computer Vision–ACCV 2014*, pages 129–145. Springer, 2015.

[92] Adrian Popescu, Eleftherios Spyromitros-Xoufis, Symeon Papadopoulos, Hervé Le Borgne, and Ioannis Kompatsiaris. Towards an automatic evaluation of retrieval performance with large scale image collections. In *MMCommons 2015 workshop*.

[93] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. The new data and new challenges in multimedia research. *CoRR*, abs/1503.01817, 2015.

[94] Quannan Li, Jiajun Wu, and Zhuowen Tu. Harvesting mid-level visual concepts from large-scale internet images. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 851–858. IEEE, 2013.

[95] Yu Su and Frédéric Jurie. Improving image classification using semantic attributes. *International Journal on Computer Vision*, 2012.

[96] Yuanqing Lin, Fengjun Lv, Shenghuo Zhu, Ming Yang, Timothee Cour, Kai Yu, Liangliang Cao, and Thomas Huang. Large-scale image classification: fast feature extraction and svm training. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1689–1696. IEEE, 2011.

[97] Yunchao Gong and Svetlana Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 817–824. IEEE, 2011.

[98] Sreemanananth Sadanand and Jason J Corso. Action bank: A high-level representation of activity in video. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1234–1241. IEEE, 2012.

[99] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3360–3367. IEEE, 2010.

[100] Tianshi Gao and Daphne Koller. Discriminative learning of relaxed hierarchy for large-scale visual recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2072–2079. IEEE, 2011.

[101] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[102] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *PAMI*, 2012.

[103] Jorge Sánchez and Florent Perronnin. High-dimensional signature compression for large-scale image classification. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1665–1672. IEEE, 2011.

[104] Jia Deng, Alexander C Berg, and Li Fei-Fei. Hierarchical semantic indexing for large scale image retrieval. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 785–792. IEEE, 2011.

[105] Alireza Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, 2009.

[106] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365–372. IEEE, 2009.

[107] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 951–958. IEEE, 2009.

[108] Vittorio Ferrari and Andrew Zisserman. Learning visual attributes. In *Advances in Neural Information Processing Systems*, pages 433–440, 2007.

[109] Keiji Yanai and Kobus Barnard. Image region entropy: a measure of visualness of web images associated with one concept. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 419–422. ACM, 2005.

[110] Marcus Rohrbach, Michael Stark, György Szarvas, Iryna Gurevych, and Bernt Schiele. What helps where–and why? semantic relatedness for knowledge transfer. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 910–917. IEEE, 2010.

[111] Ali Farhadi, Ian Endres, and Derek Hoiem. Attribute-centric recognition for cross-category generalization. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2352–2359. IEEE, 2010.

[112] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

[113] Hedi Harzallah, Frédéric Jurie, and Cordelia Schmid. Combining efficient object localization and image classification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 237–244. IEEE, 2009.

[114] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[115] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.

[116] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[117] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. What is an object? In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 73–80. IEEE, 2010.

[118] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. *International journal of computer vision*, 100(3):275–293, 2012.

[119] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Multi-fold mil training for weakly supervised object localization. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2409–2416. IEEE, 2014.

[120] Zhiyuan Shi, Timothy M Hospedales, and Tao Xiang. Bayesian joint topic modelling for weakly supervised object localisation. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2984–2991. IEEE, 2013.

[121] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell. On learning to localize objects with minimal supervision. *arXiv preprint arXiv:1403.1024*, 2014.

[122] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[123] Christian Szegedy, Alexander Toshev, and Dumitru Erhan. Deep neural networks for object detection. In *Advances in Neural Information Processing Systems*, pages 2553–2561, 2013.

[124] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. Scalable object detection using deep neural networks. In *Computer Vision and*

Pattern Recognition (CVPR), 2014 IEEE Conference on, pages 2155–2162. IEEE, 2014.

[125] Wanli Ouyang, Ping Luo, Xingyu Zeng, Shi Qiu, Yonglong Tian, Hongsheng Li, Shuo Yang, Zhe Wang, Yuanjun Xiong, Chen Qian, et al. Deepid-net: multi-stage and deformable deep convolutional neural networks for object detection. *arXiv preprint arXiv:1409.3505*, 2014.

[126] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[127] Spyros Gidaris and Nikos Komodakis. Object detection via a multi-region & semantic segmentation-aware cnn model. *arXiv preprint arXiv:1505.01749*, 2015.

[128] Christian Szegedy, Scott Reed, Dumitru Erhan, and Dragomir Anguelov. Scalable, high-quality object detection. *arXiv preprint arXiv:1412.1441*, 2014.

[129] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2014.

[130] Hong-Ming Chen and al. Sheepdog: group and tag recommendation for flickr photos by automatic search-based learning. In *ACM Multimedia 2008*.

[131] Adrian Ulges, Marcel Worring, and Thomas Breuel. Learning visual contexts for image annotation from flickr groups. *IEEE ToM*, 13(2):330–341, 2011.

[132] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, September 1975. ISSN 0001-0782. doi: 10.1145/361002.361007. URL http://doi.acm.org/10.1145/361002.361007.

[133] Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999. ISBN 020139829X.

[134] Marius Muja and David G. Lowe. Scalable nearest neighbor algorithms for high dimensional data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(11):2227–2240, 2014. doi: 10.1109/TPAMI.2014.2321376. URL http://doi.ieeecomputersociety.org/10.1109/TPAMI.2014.2321376.

[135] Sunil Arya, David M. Mount, Nathan S. Netanyahu, Ruth Silverman, and Angela Y. Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *J. ACM*, 45(6):891–923, November 1998. ISSN 0004-5411. doi: 10.1145/293347.293348. URL http://doi.acm.org/10.1145/293347.293348.

[136] Jeffrey S. Beis and David G. Lowe. Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, CVPR '97, pages 1000–, Washington, DC, USA, 1997. IEEE Computer Society. ISBN 0-8186-7822-4. URL http://dl.acm.org/citation.cfm?id=794189.794431.

[137] Mohamed Aly, Mario Munich, and Pietro Perona. Distributed Kd-Trees for Retrieval from Very Large Image Collections. In *BMVC 2011*.

[138] Luiz André Barroso, Jeffrey Dean, and Urs Hölzle. Web search for a planet: The google cluster architecture. *Micro, Ieee*, 23(2):22–28, 2003.

[139] Yang Cao, Changhu Wang, Zhiwei Li, Liqing Zhang, and Lei Zhang. Spatial-bag-of-features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3352–3359. IEEE, 2010.

[140] David M Chen, Sam S Tsai, Vijay Chandrasekhar, Gabriel Takacs, Ramakrishna Vedantham, Radek Grzeszczuk, and Bernd Girod. Inverted index compression for scalable image matching. In *DCC*, page 525, 2010.

[141] Artem Babenko and Victor Lempitsky. The inverted multi-index. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3069–3076. IEEE, 2012.

[142] Liang Zheng, Shengjin Wang, Ziqiong Liu, and Qi Tian. Packing and padding: Coupled multi-index for accurate image retrieval. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1947–1954. IEEE, 2014.

[143] Romain Tavenard, Hervé Jégou, and Laurent Amsaleg. Balancing clusters to reduce response time variability in large scale image search. In *Content-Based Multimedia Indexing (CBMI), 2011 9th International Workshop on*, pages 19–24. IEEE, 2011.

[144] Ruoyu Liu, Yao Zhao, Shikui Wei, Zhenfeng Zhu, Lixin Liao, and Shuang Qiu. Indexing of cnn features for large scale image search. *arXiv preprint arXiv:1508.00217*, 2015.

[145] BJ Fogg and Hsiang Tseng. The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit*, pages 80–87. ACM, 1999.

[146] Andreas Juffinger, Michael Granitzer, and Elisabeth Lex. Blog credibility ranking by exploiting verified content. In *Proceedings of the 3rd workshop on Information credibility on the web*, pages 51–58. ACM, 2009.

[147] Wouter Weerkamp and Maarten De Rijke. Credibility improves topical blog post retrieval. In *46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2008.

[148] Jiang Bian, Yandong Liu, Ding Zhou, Eugene Agichtein, and Hongyuan Zha. Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In *Proceedings of the 18th international conference on World wide web*, pages 51–60. ACM, 2009.

[149] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. Finding high-quality content in social media. In *Proceedings of the international conference on Web search and web data mining*, pages 183–194. ACM, 2008.

[150] Irma Becerra-Fernandez. Facilitating the online search of experts at nasa using expert seeker people-finder. In *Proceedings of the 3rd International Conference on Practical Aspects of Knowledge Management (PAKM), Basel, Switzerland*, 2000.

[151] Ahmad Kardan, Mehdi Garakani, and Bamdad Bahrani. A method to automatically construct a user knowledge model in a forum environment. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 717–718. ACM, 2010.

[152] Jun Zhang, Mark S Ackerman, and Lada Adamic. Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web*, pages 221–230. ACM, 2007.

[153] Krisztian Balog, Maarten De Rijke, and Wouter Weerkamp. Bloggers as experts: feed distillation using expert retrieval models. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 753–754. ACM, 2008.

[154] Michael G Noll, Ching-man Au Yeung, Nicholas Gibbins, Christoph Meinel, and Nigel Shadbolt. Telling experts from spammers: expertise ranking in folksonomies. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 612–619. ACM, 2009.

[155] Pranam Kolari, Tim Finin, Kelly Lyons, and Yelena Yesha. Expert search using internal corporate blogs. In *Proceedings of SIGIR Workshop: Future Challenges in Expertise Retrieval*, pages 7–10, 2008.

[156] Einat Amitay, David Carmel, Nadav Har'El, Shila Ofek-Koifman, Aya Soffer, Sivan Yogev, and Nadav Golbandi. Social search and discovery using a unified approach.

In *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, pages 199–208. ACM, 2009.

[157] Ido Guy, Uri Avraham, David Carmel, Sigalit Ur, Michal Jacovi, and Inbal Ronen. Mining expertise and interests from social media. In *Proceedings of the 21st international conference on World Wide Web*, 2013.

[158] Pawel Jurczyk and Eugene Agichtein. Discovering authorities in question answer communities by using link analysis. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 919–922. ACM, 2007.

[159] Pawel Jurczyk and Eugene Agichtein. Hits on question answer portals: exploration of link analysis for author ranking. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 845–846. ACM, 2007.

[160] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[161] Xiaoyong Liu, W Bruce Croft, and Matthew Koll. Finding experts in community-based question-answering services. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 315–316. ACM, 2005.

[162] Jan Rybak, Krisztian Balog, and Kjetil Nørvåg. Temporal expertise profiling. In *Advances in Information Retrieval*, pages 540–546. Springer, 2014.

[163] R David Lankes. Trusting the internet: new approaches to credibility tools. *The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning*, pages 101–121, 2007.

[164] Thomas J Johnson and Barbara K Kaye. In blog we trust? deciphering credibility of components of the internet among politically interested internet users. *Computers in Human Behavior*, 25(1):175–182, 2009.

[165] Teun Lucassen and Jan Maarten Schraagen. Trust in wikipedia: how users trust information from an unknown source. In *Proceedings of the 4th workshop on Information credibility*, pages 19–26. ACM, 2010.

[166] Robert Thomson, Naoya Ito, Hinako Suda, Fangyu Lin, Yafei Liu, Ryo Hayasaka, Ryuzo Isochi, and Zian Wang. Trusting tweets: The fukushima disaster and information source credibility on twitter. In *Proceedings of the 9th International Conference on Information Systems for Crisis Response and Management*, 2012.

[167] Grace YoungJoo Jeon and Soo Young Rieh. Do you trust answers?: Credibility judgments in social search using social q&a sites. *social networks*, 2:14, 2013.

[168] Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. Combating web spam with trustrank. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 576–587. VLDB Endowment, 2004.

[169] Catalina Laura Toma. Counting on friends: Cues to perceived trustworthiness in facebook profiles. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[170] Ramanthan Guha, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Propagation of trust and distrust. In *Proceedings of the 13th international conference on World Wide Web*, pages 403–412. ACM, 2004.

[171] Giacomo Bachi, Michele Coscia, Anna Monreale, and Fosca Giannotti. Classifying trust/distrust relationships in online social networks. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 552–557. IEEE, 2012.

[172] Cai-Nicolas Ziegler and Georg Lausen. Propagation models for trust and distrust in social networks. *Information Systems Frontiers*, 7(4-5):337–358, 2005.

[173] Brian Hilligoss and Soo Young Rieh. Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing & Management*, 44(4):1467–1484, 2008.

[174] Soo Young Rieh and Nicholas J Belkin. Understanding judgment of information quality and cognitive authority in the www. In *Proceedings of the 61st annual meeting of the american society for information science*, volume 35, pages 279–289. Citeseer, 1998.

[175] Miriam J Metzger. Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, 58(13):2078–2091, 2007.

[176] Omar Alonso, Chad Carson, David Gerster, Xiang Ji, and Shubha U Nabar. Detecting uninteresting content in text streams. In *SIGIR Crowdsourcing for Search Evaluation Workshop*, 2010.

[177] Jill Burstein and Magdalena Wolska. Toward evaluation of writing style: finding overly repetitive word use in student essays. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 35–42. Association for Computational Linguistics, 2003.

[178] Martin Chodorow and Claudia Leacock. An unsupervised method for detecting grammatical errors. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 140–147. Association for Computational Linguistics, 2000.

[179] Lawrence M Rudner and Tahung Liang. Automated essay scoring using bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2), 2002.

[180] Yigal Attali and Jill Burstein. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3), 2006.

[181] Luo Si and Jamie Callan. A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576. ACM, 2001.

[182] Kevyn Collins-Thompson and Jamie Callan. A language modeling approach to predicting reading difficulty. In *Proceedings of HLT/NAACL*, volume 4, 2004.

[183] Jamie Callan and Maxine Eskenazi. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of NAACL HLT*, pages 460–467, 2007.

[184] Sarah E Petersen and Mari Ostendorf. A machine learning approach to reading level assessment. *Computer Speech & Language*, 23(1):89–106, 2009.

[185] Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195. Association for Computational Linguistics, 2008.

[186] Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 229–237. Association for Computational Linguistics, 2009.

[187] Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284. Association for Computational Linguistics, 2010.

[188] G Harry McLaughlin. Smog grading: A new readability formula. *Journal of reading*, 12(8):639–646, 1969.

[189] Susumu Akamine, Daisuke Kawahara, Yoshikiyo Kato, Tetsuji Nakagawa, Yutaka I Leon-Suematsu, Takuya Kawada, Kentaro Inui, Sadao Kurohashi, and Yutaka Kidawara. Organizing information on the web to support user judgments on information credibility. In *Universal Communication Symposium (IUCS), 2010 4th International*, pages 123–130. IEEE, 2010.

[190] Francis Heylighen and Jean-Marc Dewaele. Variation in the contextuality of language: An empirical measure. *Foundations of Science*, 7(3):293–340, 2002.

[191] Isabel Drost and Tobias Scheffer. Thwarting the nigritude ultramarine: Learning to identify link spam. In *Machine Learning: ECML 2005*, pages 96–107. Springer, 2005.

[192] Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. Detecting spam web pages through content analysis. In *Proceedings of the 15th international conference on World Wide Web*, pages 83–92. ACM, 2006.

[193] Ricardo Baeza-Yates, Carlos Castillo, Vicente López, and Cátedra Telefónica. Pagerank increase under different collusion topologies. In *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, pages 17–24, 2005.

[194] Luca Becchetti, Carlos Castillo, Debora Donato, Ricardo Baeza-Yates, and Stefano Leonardi. Link analysis for web spam detection. *ACM Transactions on the Web (TWEB)*, 2(1):2, 2008.

[195] Reid Andersen, Christian Borgs, Jennifer Chayes, John Hopcroft, Kamal Jain, Vahab Mirrokni, and Shanghua Teng. Robust pagerank and locally computable spam detection features. In *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pages 69–76. ACM, 2008.

[196] Adam Thomason. Blog spam: A review. In *Conference on Email and Anti-Spam (CEAS)*, 2007.

[197] Fabrício Benevenuto, Tiago Rodrigues, Virgilio Almeida, Jussara Almeida, and Keith Ross. Video interactions in online video social networks. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 5 (4):30, 2009.

[198] Fabrício Benevenuto, Tiago Rodrigues, Adriano Veloso, Jussara Almeida, Marcos Gonçalves, and Virgílio Almeida. Practical detection of spammers and content promoters in online video sharing systems. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(3):688–701, 2012.

[199] Fabrıcio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgılio Almeida. Detecting spammers on twitter. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, volume 6, 2010.

[200] Kyumin Lee, James Caverlee, and Steve Webb. Uncovering social spammers: social honeypots+ machine learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 435–442. ACM, 2010.

[201] Hongyu Gao, Jun Hu, Christo Wilson, Zhichun Li, Yan Chen, and Ben Y Zhao. Detecting and characterizing social spam campaigns. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 35–47. ACM, 2010.

[202] Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the international conference on Web search and web data mining*, pages 219–230. ACM, 2008.

[203] Gordon V Cormack. Email spam filtering: A systematic review. *Foundations and Trends in Information Retrieval*, 1(4):335–455, 2007.

[204] Ching-Tung Wu, Kwang-Ting Cheng, Qiang Zhu, and Yi-Leh Wu. Using visual features for anti-spam filtering. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 3, pages III–509. IEEE, 2005.

[205] Fabrício Benevenuto, Tiago Rodrigues, Virgílio Almeida, Jussara Almeida, and Marcos Gonçalves. Detecting spammers and content promoters in online video social networks. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 620–627. ACM, 2009.

[206] Yunjie Calvin Xu and Zhiwei Chen. Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology*, 57(7):961–973, 2006.

[207] Soo Young Rieh. Judgment of information quality and cognitive authority in the web. *Journal of the American Society for Information Science and Technology*, 53 (2):145–161, 2002.

[208] Eric Nichols, Koji Murakami, Kentaro Inui, and Yuji Matsumoto. Constructing a scientific blog corpus for information credibility analysis. In *Proc. of the Annual Meeting of ANLP*, 2009.

[209] Luis Sanz, Héctor Allende, and Marcelo Mendoza. Text content reliability estimation in web documents: a new proposal. *Computational Linguistics and Intelligent Text Processing*, pages 438–449, 2012.

[210] Parikshit Sondhi, V Vydiswaran, and ChengXiang Zhai. Reliability prediction of webpages in the medical domain. *Advances in Information Retrieval*, pages 219–231, 2012.

[211] Farah Alsudani and Matthew Casey. The effect of aesthetics on web credibility. In *Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology*, BCS-HCI '09, pages 512–519, Swinton, UK, UK, 2009. British Computer Society. URL http://dl.acm.org/citation.cfm?id=1671011.1671077.

[212] Andrew J Flanagin and Miriam J Metzger. The perceived credibility of personal web page information as influenced by the sex of the source. *Computers in Human Behavior*, 19(6):683–701, 2003.

[213] Cory L Armstrong and Melinda J McAdams. Blogs of information: How gender cues and individual motivations influence perceptions of credibility. *Journal of Computer-Mediated Communication*, 14(3):435–456, 2009.

[214] Simon Duncan and Birgit Pfau-Effinge, editors. *Gender, Economy and Culture in the European Union*. Routledge Research in Gender and Society, 2012.

[215] R. E. Petty and J. T. Cacioppo. The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, 19, 1986.

[216] S. Chaiken. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, 39(5), 1980.

[217] Radoslaw Nielek, Aleksander Wawer, Michal Jankowski-Lorek, and Adam Wierzbicki. Temporal, cultural and thematic aspects of web credibility. In *Social Informatics*, pages 419–428. Springer, 2013.

[218] Chuan Luo, Xin Robert Luo, Laurie Schatzberg, and Choon Ling Sia. Impact of informational factors on online recommendation credibility: The moderating role of source credibility. *Decision Support Systems*, 2013.

[219] Qian Xu. Should i trust him? the effects of reviewer profile characteristics on ewom credibility. *Computers in Human Behavior*, 33:136–144, 2014.

[220] Heelye Park, Zheng Xiang, Bharath Josiam, and Haejung Kim. Personal profile information as cues of credibility in online travel reviews. *Anatolia*, 25(1):13–23, 2014.

[221] G.L. Patzer. Source credibility as a function of communicator physical attractiveness. *Journal of Business Research*, 11(2), 1983.

[222] Paul Benjamin Lowry, David W Wilson, and William L Haig. A picture is worth a thousand words: source credibility theory applied to logo and website design for heightened credibility and consumer trust. *International Journal of Human-Computer Interaction*, 30(1):63–93, 2014.

[223] Tristan Endsley, Yu Wu, and James Reep. The source of the story: Evaluating the credibility of crisis information sources. *Proceedings of the Information Systems for Crisis Response and Management (ISCRAM)*, 1:158–162, 2014.

[224] Rui Lopes and Luis Carriço. On the credibility of wikipedia: an accessibility perspective. In *Proceedings of the 2nd ACM workshop on Information credibility on the web*, pages 27–34. ACM, 2008.

[225] Julian K Ayeh, Norman Au, and Rob Law. "do we believe in tripadvisor?" examining credibility perceptions and online travelers' attitude toward using user-generated content. *Journal of Travel Research*, page 0047287512475217, 2013.

[226] Hui Jimmy Xie, Li Miao, Pei-Jou Kuo, and Bo-Youn Lee. Consumers' responses to ambivalent online hotel reviews: The role of perceived source credibility and pre-decisional disposition. *International Journal of Hospitality Management*, 30 (1):178–183, 2011.

[227] T. S. Robertson and J. R. Rossiter. Children and commercial persuasion: An attribution theory analysis. *Journal of Consumer Research*, 1(1), 1974.

[228] B. J. Fogg, L. Marable, J. Stanford, and E. R. Tauber. How do people evaluate a web site's credibility? Technical report, The Stanford Persuasive Technology Lab, 2002.

[229] Wei Zha and H Denis Wu. The impact of online disruptive ads on users' comprehension, evaluation of site credibility, and sentiment of intrusiveness. *American Communication Journal*, 16(2), 2014.

[230] Choicestream. Choicestream survey: Consumer opinion on online advertising and audience targeting, 2013.

[231] Aditi Gupta and Ponnurangam Kumaraguru. Credibility ranking of tweets during high impact events. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, page 2. ACM, 2012.

[232] Jeff Pasternack and Dan Roth. Latent credibility analysis. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1009–1020. International World Wide Web Conferences Steering Committee, 2013.

[233] Aditya Pal and Scott Counts. Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 45–54. ACM, 2011.

[234] Thanasis G Papaioannou, Karl Aberer, Katarzyna Abramczuk, Paulina Adamska, and Adam Wierzbicki. Game-theoretic models of web credibility. In *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, pages 27–34. ACM, 2012.

[235] Wojciech Jaworski, Emilia Rejmund, and Adam Wierzbicki. Credibility microscope: relating web page credibility evaluations to their textual content. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, volume 1, pages 297–302. IEEE, 2014.

[236] Aleksander Wawer, Radoslaw Nielek, and Adam Wierzbicki. Predicting webpage credibility using linguistic features. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 1135–1140, Republic and Canton of Geneva, Switzerland, 2014. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-2745-9. doi: 10.1145/2567948.2579000. URL http://dx.doi.org/10.1145/2567948.2579000.

[237] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Predicting information credibility in time-sensitive social media. *Internet Research*, 23(5):560–588, 2013.

[238] Xin Liu, Radoslaw Nielek, Adam Wierzbicki, and Karl Aberer. Defending imitating attacks in web credibility evaluation systems. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 1115–1122, 2013.

[239] Wouter Weerkamp and Maarten de Rijke. Credibility-inspired ranking for blog post retrieval. *Information retrieval*, pages 1–35, 2012.

[240] Lada A Adamic, Jun Zhang, Eytan Bakshy, and Mark S Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web*, pages 665–674. ACM, 2008.

[241] Yukiko Kawai, Yusuke Fujita, Tadahiko Kumamoto, Jianwei Jianwei, and Katsumi Tanaka. Using a sentiment map for visualizing credibility of news sites on the web. In *Proceedings of the 2nd ACM workshop on Information credibility on the web*, pages 53–58. ACM, 2008.

[242] Koji Murakami, Eric Nichols, Suguru Matsuyoshi, Asuka Sumida, Shouko Masuda, Kentaro Inui, and Yuji Matumoto. Statement map: assisting information crediblity analysis by visualizing arguments. In *Proceedings of the 3rd workshop on Information credibility on the web*, pages 43–50. ACM, 2009.

[243] John ODonovan, Byungkyu Kang, Greg Meyer, Tobias Hollerer, and Sibel Adalii. Credibility in context: An analysis of feature distributions in twitter. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 293–301. IEEE, 2012.

[244] Andrew J Flanagin, Miriam J Metzger, Rebekah Pure, Alex Markov, and Ethan Hartsell. Mitigating risk in ecommerce transactions: perceptions of information credibility and the role of user-generated ratings in product quality and purchase intention. *Electronic Commerce Research*, 14(1):1–23, 2014.

[245] Maria Rafalak, Katarzyna Abramczuk, and Adam Wierzbicki. Incredible: Is (almost) all web content trustworthy? analysis of psychological factors related to website credibility evaluation. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 1117–1122. International World Wide Web Conferences Steering Committee, 2014.

[246] T. J. Johnson and D. Perlmutter. The facebook election. *Mass Communication and Society*, 2010.

[247] A. Geiber. Digital divas: Women, politics and the social network. Technical Report D-63, Cambridge, MA: Joan Shorenstein Center on the Press, 2011.

[248] Thomas J Johnson and Barbara K Kaye. Credibility of social network sites for political information among politically interested internet users. *Journal of Computer-Mediated Communication*, 19(4):957–974, 2014.

[249] Mohammad-Ali Abbasi and Huan Liu. Measuring user credibility in social media. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 441–448. Springer, 2013.

[250] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P Gummadi. Measuring user influence in twitter: The million follower fallacy. In *4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, page 8, 2010.

[251] Michael P O'Mahony and Barry Smyth. Learning to recommend helpful hotel reviews. In *Proceedings of the third ACM conference on Recommender systems*, pages 305–308. ACM, 2009.

[252] Michael P O'Mahony and Barry Smyth. Using readability tests to predict helpful product reviews. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, pages 164–167. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2010.

[253] Chad Edwards, Patric R Spence, Christina J Gentile, America Edwards, and Autumn Edwards. How much klout do you have... a test of system generated cues on source credibility. *Computers in Human Behavior*, 29(5):A12–A16, 2013.

[254] Azizul Yaakop, Marhana Mohamed Anuar, and Khatijah Omar. Like it or not: issue of credibility in facebook advertising. *Asian Social Science*, 9(3):p154, 2013.

[255] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. URL http://ilpubs.stanford.edu:8090/422/.

[256] Christopher S Campbell, Paul P Maglio, Alex Cozzi, and Byron Dom. Expertise identification using email communications. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 528–531. ACM, 2003.

[257] Byron Dom, Iris Eiron, Alex Cozzi, and Yi Zhang. Graph-based ranking algorithms for e-mail expertise analysis. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 42–48. ACM, 2003.

[258] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.

[259] David Westerman, Patric R Spence, and Brandon Van Der Heide. A social network as information: The effect of system generated reports of connectedness on credibility on twitter. *Computers in Human Behavior*, 28(1):199–206, 2012.

[260] Jiang Yang, Scott Counts, Meredith Ringel Morris, and Aaron Hoff. Microblog credibility perceptions: comparing the usa and china. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 575–586. ACM, 2013.

[261] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 441–450. ACM, 2012.

[262] Byungkyu Kang, John O'Donovan, and Tobias Höllerer. Modeling topic specific credibility on twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 179–188. ACM, 2012.

[263] Manish Gupta, Peixiang Zhao, and Jiawei Han. Evaluating event credibility on twitter. *Proc. of SIAM*, 2012.

[264] Marie Truelove, Maria Vasardani, and Stephan Winter. Towards credibility of micro-blogs: characterising witness accounts. *GeoJournal*, pages 1–21, 2014.

[265] Claudia Wagner, Vera Liao, Peter Pirolli, Les Nelson, and Markus Strohmaier. It's not in their tweets: Modeling topical expertise of twitter users. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 91–100. IEEE, 2012.

[266] Naveen Kumar Sharma, Saptarshi Ghosh, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. Inferring who-is-who in the twitter social network. In *Proceedings of the 2012 ACM workshop on Workshop on online social networks*, pages 55–60. ACM, 2012.

[267] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we rt? In *Proceedings of the First Workshop on Social Media Analytics*, pages 71–79. ACM, 2010.

[268] Martin Hentschel, Omar Alonso, Scott Counts, and Vasileios Kandylas. Finding users we trust: Scaling up verified twitter users using their communication patterns. In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014.*, 2014. URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8063.

[269] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*, pages 249–252. ACM, 2011.

[270] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. Detecting and tracking the spread of astroturf memes in microblog streams. *arXiv preprint arXiv:1011.3768*, 2010.

[271] Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. @ spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security*, pages 27–37. ACM, 2010.

[272] Kamran Massoudi, Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp. Incorporating query expansion and quality indicators in searching microblog posts. *Advances in Information Retrieval*, pages 362–367, 2011.

[273] Sujoy Sikdar, Byungkyu Kang, John O'Donovan, Tobias Hollerer, and Sibel Adah. Understanding information credibility on twitter. In *Social Computing (SocialCom), 2013 International Conference on*, pages 19–24. IEEE, 2013.

[274] S Sikdar, S Adali, M Amin, T Abdelzaher, K Chan, J-H Cho, B Kang, and J O'Donovan. Finding true and credible information on twitter. In *Information Fusion (FUSION), 2014 17th International Conference on*, pages 1–8. IEEE, 2014.

[275] David Westerman, Patric R Spence, and Brandon Van Der Heide. Social media as information source: Recency of updates and credibility of information. *Journal of Computer-Mediated Communication*, 19(2):171–183, 2014.

[276] Suliman Aladhadh, Xiuzhen Zhang, and Mark Sanderson. Tweet author location impacts on tweet credibility. In *Proceedings of the 2014 Australasian Document Computing Symposium*, page 73. ACM, 2014.

[277] Shafiza Mohd Shariff, Xiuzhen Zhang, and Mark Sanderson. User perception of information credibility of news on twitter. In *Advances in Information Retrieval*, pages 513–518. Springer, 2014.

[278] Petros Kostagiolas, Nikolaos Korfiatis, Panos Kourouthanasis, and Georgios Alexias. Work-related factors influencing doctors search behaviors and trust toward medical information resources. *International Journal of Information Management*, 34(2):80–88, 2014.

[279] Qi Su, Dmitry Pavlov, Jyh-Herng Chow, and Wendell C Baker. Internet-scale collection of human-reviewed data. In *Proceedings of the 16th international conference on World Wide Web*, pages 231–240. ACM, 2007.

[280] Jiwoon Jeon, W Bruce Croft, Joon Ho Lee, and Soyeon Park. A framework to predict the quality of answers with non-textual features. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 228–235. ACM, 2006.

[281] Qi Su, Chu-Ren Huang, and Helen Kai-yun Chen. Evidentiality for text trustworthiness detection. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, pages 10–17. Association for Computational Linguistics, 2010.

[282] Mohamed Bouguessa, Benoît Dumoulin, and Shengrui Wang. Identifying authoritative actors in question-answering forums: The case of yahoo! answers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 866–874, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-193-4. doi: 10.1145/1401890.1401994. URL http://doi.acm.org/10.1145/1401890.1401994.

[283] Duen-Ren Liu, Yu-Hsuan Chen, Wei-Chen Kao, and Hsiu-Wen Wang. Integrating expert profile, reputation and link analysis for expert finding in question-answering websites. *Information Processing & Management*, 49(1):312 – 329, 2013.

[284] Jiang Bian, Yandong Liu, Ding Zhou, Eugene Agichtein, and Hongyuan Zha. Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 51–60, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-487-4. doi: 10.1145/1526709.1526717. URL http://doi.acm.org/10.1145/1526709.1526717.

[285] Qiaoling Liu and Eugene Agichtein. Modeling answerer behavior in collaborative question answering systems. In *Proc. of ECIR'11*, pages 67–79, 2011.

[286] Aditya Pal and Joseph A. Konstan. Expert identification in community question answering: exploring question selection bias. In *Proc. of ACM CIKM'10*, pages 1505–1508, 2010. ISBN 978-1-4503-0099-5.

[287] Alexandru Lucian Ginsca, Adrian Popescu, Bogdan Ionescu, Anil Armagan, and Ioannis Kanellos. Toward an estimation of user tagging credibility for social image retrieval. In *Proceedings of the ACM International Conference on Multimedia*, pages 1021–1024. ACM, 2014.

[288] Rodrigo T Calumby, Vinícius P Santana, Felipe S Cordeiro, Otávio AB Penatti, Lin T Li, Giovani Chiachia, and Ricardo da S Torres. Recod@ mediaeval 2014: Diverse social images retrieval. *Working Notes of MediaEval*, 2014.

[289] Duc-Tien Dang-Nguyen, Luca Piras, Giorgio Giacinto, Giulia Boato, and FGB De Natale. Retrieval of diverse images by pre-filtering and hierarchical clustering. *Working Notes of MediaEval*, 2014.

[290] Bogdan Ionescu, Adrian Popescu, Mihai Lupu, Alexandru Lucian Gînscă, Bogdan Boteanu, and Henning Müller. Div150cred: A social image retrieval result diversification with user tagging credibility dataset. In *Proceedings of the 6th ACM Multimedia Systems Conference*, MMSys '15, pages 207–212, New York, NY,

USA, 2015. ACM. ISBN 978-1-4503-3351-1. doi: 10.1145/2713168.2713192. URL http://doi.acm.org/10.1145/2713168.2713192.

[291] Kristina Lerman. Social information processing in news aggregation. *Internet Computing, IEEE*, 11(6):16–28, 2007.

[292] Nicholas Diakopoulos, Sergio Goldenberg, and Irfan Essa. Videolyzer: quality analysis of online informational video for bloggers and journalists. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 799–808. ACM, 2009.

[293] Nicholas Diakopoulos and Irfan Essa. An annotation model for making sense of information quality in online video. In *Proceedings of the 3rd International Conference on the Pragmatic Web: Innovating the Interactive Society*, pages 31–34. ACM, 2008.

[294] Derek O'Callaghan, Martin Harrigan, Joe Carthy, and Pádraig Cunningham. Network analysis of recurring youtube spam campaigns. In *Procs. of the 6th Intl. AAAI Conference on Weblogs and Social Media (ICWSM 12)*, 2012.

[295] Susanne Boll. Multitube–where web 2.0 and multimedia could meet. *Multimedia, IEEE*, 14(1):9–13, 2007.

[296] Vlad Bulakh, Christopher W Dunn, and Minaxi Gupta. Identifying fraudulently promoted online videos. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 1111–1116. International World Wide Web Conferences Steering Committee, 2014.

[297] Ling Xu, Qiang Ma, and Masatoshi Yoshikawa. Credibility-oriented ranking of multimedia news based on a material-opinion model. *Web-Age Information Management*, pages 290–301, 2011.

[298] Ling Xu, Qiang Ma, and Masatoshi Yoshikawa. A cross-media method of stakeholder extraction for news contents analysis. In *Web-Age Information Management*, pages 232–237. Springer, 2010.

[299] Yusuke Yamamoto and Katsumi Tanaka. Imagealert: credibility analysis of text-image pairs on the web. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, pages 1724–1731. ACM, 2011.

[300] Manos Tsagkias, Martha Larson, and Maarten De Rijke. Predicting podcast preference: An analysis framework and its application. *Journal of the American Society for Information Science and Technology*, 61(2):374–391, 2009.

[301] Soo Young Rieh and David R Danielson. Credibility: A multidisciplinary framework. *Annual review of information science and technology*, 41(1):307–364, 2007.

[302] Shawn Tseng and BJ Fogg. Credibility and computing technology. *Communications of the ACM*, 42(5):39–44, 1999.

[303] Craig Macdonald and Iadh Ounis. The trec blogs06 collection: Creating and analysing a blog test collection. *Department of Computer Science, University of Glasgow Tech Report TR-2006-224*, 1:3–1, 2006.

[304] Haifeng Liu, Ee-Peng Lim, Hady W Lauw, Minh-Tam Le, Aixin Sun, Jaideep Srivastava, and Young Kim. Predicting trusts among users of online communities: an epinions case study. In *Proceedings of the 9th ACM conference on Electronic commerce*, pages 310–319. ACM, 2008.

[305] Paolo Massa and Paolo Avesani. Controversial users demand local trust metrics: An experimental study on epinions. com community. In *Proceedings of the National Conference on artificial Intelligence*, volume 20, page 121. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

[306] Gabriel De La Calzada and Alex Dekhtyar. On measuring the quality of wikipedia articles. In *Proceedings of the 4th workshop on Information credibility*, pages 11–18. ACM, 2010.

[307] Yu Suzuki and Masatoshi Yoshikawa. Qualityrank: assessing quality of wikipedia articles by mutually evaluating editors and texts. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, pages 307–308. ACM, 2012.

[308] Jennifer Rowley and Frances Johnson. Understanding trust formation in digital information sources: The case of wikipedia. *J. Inf. Sci.*, 39(4):494–508, August 2013. ISSN 0165-5515. doi: 10.1177/0165551513477820. URL http://dx.doi.org/10.1177/0165551513477820.

[309] Peter Pirolli, Evelin Wollny, and Bongwon Suh. So you know you're getting the best possible information: a tool that increases wikipedia credibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1505–1508. ACM, 2009.

[310] Dan Pelleg, Elad Yom-Tov, and Yoelle Maarek. Can you believe an anonymous contributor? on truthfulness in yahoo! answers. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 411–420. IEEE, 2012.

[311] Wei-Chen Kao, Duen-Ren Liu, and Shiu-Wen Wang. Expert finding in question-answering websites: a novel hybrid approach. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 867–871. ACM, 2010.

[312] Aditya Pal, Shuo Chang, and Joseph A Konstan. Evolution of experts in question answering communities. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 274–281, 2012.

[313] Benjamin V Hanrahan, Gregorio Convertino, and Les Nelson. Modeling problem difficulty and expertise in stackoverflow. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*, pages 91–94. ACM, 2012.

[314] Yin Aphinyanaphongs, Constantin Aliferis, et al. Text categorization models for identifying unproven cancer treatments on the web. *Studies in health technology and informatics*, 129(2):968, 2007.

[315] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, July 2006. ISSN 0899-7667. doi: 10.1162/neco.2006.18.7.1527.

[316] Lyndon S. Kennedy and al. To search or to label?: predicting the performance of search-based automatic image classifiers. In *ACM MIR 2006*.

[317] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[318] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. In *CVPR*. IEEE, 2014.

[319] Jeff Donahue and al. Decaf: A deep convolutional activation feature for generic visual recognition. *CoRR*, 2013.

[320] Zheng Song, Qiang Chen, Zhongyang Huang, Yang Hua, and Shuicheng Yan. Contextualizing object detection and classification. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1585–1592. IEEE, 2011.

[321] Yunchao Wei, Wei Xia, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. Cnn: Single-label to multi-label. *arXiv preprint arXiv:1406.5726*, 2014.

[322] Hao Yang, Joey Tianyi Zhou, Yu Zhang, Bin-Bin Gao, Jianxin Wu, and Jianfei Cai. Can partial strong labels boost multi-label object recognition? *arXiv preprint arXiv:1504.05843*, 2015.

[323] Jianwei Luo, Jianguo Li, Jun Wang, Zhiguo Jiang, and Yurong Chen. Deep attributes from context-aware regional neural codes. *arXiv preprint arXiv:1509.02470*, 2015.

[324] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014.

[325] Kevin Lin, Huei-Fang Yang, Jen-Hao Hsiao, and Chu-Song Chen. Deep learning of binary hash codes for fast image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 27–35, 2015.

[326] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, pages 1–42, 2014.

[327] Rong-En Fan and al. Liblinear: A library for large linear classification. *JMLR*, 2008.

[328] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

[329] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[330] Xirong Li and Cees GM Snoek. Classifying tag relevance with relevant positive and negative examples. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 485–488. ACM, 2013.

[331] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

[332] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[333] Thomas Deselaers and Vittorio Ferrari. Visual and semantic similarity in imagenet. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1777–1784. IEEE, 2011.

[334] Jay J Jiang and David W Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*, 1997.

[335] Radu Andrei Negoescu and Daniel Gatica-Perez. Analyzing flickr groups. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 417–426. ACM, 2008.

[336] Dongyuan Lu and Qiudan Li. Exploiting semantic hierarchies for flickr group. In *Active Media Technology*, pages 74–85. Springer, 2010.

[337] Shiai Zhu, Chong-Wah Ngo, and Yu-Gang Jiang. Sampling and ontologically pooling web images for visual concept learning. *Multimedia, IEEE Transactions on*, 14 (4):1068–1078, 2012.

[338] Elisavet Chatzilari. Using tagged images of low visual ambiguity to boost the learning efficiency of object detectors. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 1027–1030. ACM, 2013.

[339] Lyndon S. Kennedy and Mor Naaman. Generating diverse and representative image search results for landmarks. In *Proc. of WWW 2008*, pages 297–306.

[340] Arnold W. M. Smeulders and al. Content-based image retrieval at the end of the early years. *IEEE PAMI*, 22(12):1349–1380, December 2000. ISSN 0162-8828. doi: 10.1109/34.895972. URL http://dx.doi.org/10.1109/34.895972.

[341] D. Zhang and al. A review on automatic image annotation techniques. *Patt. Recognition*, 45(1), 2012. ISSN 0031-3203.

[342] M. Villegas and R. Paredes. Overview of the ImageCLEF 2014 Scalable Concept Image Annotation Task. In *CLEF 2014 Working Notes*.

[343] Lingqiao Liu, Lei Wang, and Xinwang Liu. In defense of soft-assignment coding. In *ICCV*. IEEE, 2011.

[344] Kevin Beyer and al. When is nearest neighbor meaningful? In *ICDT*. Springer, 1999.

[345] Kai Yu, Tong Zhang, and Yihong Gong. Nonlinear learning using local coordinate coding. In *NIPS*, 2009.

[346] Jinjun Wang and al. Locality-constrained linear coding for image classification. In *CVPR*. IEEE, 2010.

[347] Ondrej Chum, James Philbin, Andrew Zisserman, et al. Near duplicate image detection: min-hash and tf-idf weighting. In *BMVC*, volume 810, pages 812–815, 2008.

[348] Theodora Tsikrika and al. Building reliable and reusable test collections for image retrieval: The wikipedia task at imageclef. *IEEE MultiMedia*, 2012.

[349] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 39–43. ACM, 2008.

[350] Christian Beecks, Merih Seran Uysal, and Thomas Seidl. A comparative study of similarity measures for content-based multimedia retrieval. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1552–1557. IEEE, 2010.

[351] Wei Wang, Beng Chin Ooi, Xiaoyan Yang, Dongxiang Zhang, and Yueting Zhuang. Effective multi-modal retrieval based on stacked auto-encoders. *Proceedings of the VLDB Endowment*, 7(8):649–660, 2014.

[352] Jinhui Tang, Zechao Li, Liyan Zhang, and Qingming Huang. Semantic-aware hashing for social image retrieval. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 483–486. ACM, 2015.

[353] M. Everingham and al. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[354] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright, editors, *Optimization for Machine Learning*, pages 351–368. MIT Press, 2011. URL http://leon.bottou.org/papers/bottou-bousquet-2011.

[355] Xirong Li, Cees GM Snoek, Marcel Worring, Dennis Koelma, and Arnold WM Smeulders. Bootstrapping visual categorization with relevant negatives. *Multimedia, IEEE Transactions on*, 15(4):933–945, 2013.

[356] Morgan Ames and Mor Naaman. Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 971–980. ACM, 2007.

[357] Jack Kustanowitz and Ben Shneiderman. Motivating annotation for personal digital photo libraries: Lowering barriers while raising incentives.

[358] Chen Ye and Oded Nov. Exploring user contributed information in social computing systems: quantity versus quality. *Online Information Review*, 37(5):752–770, 2013.

[359] Bogdan Ionescu, Anca-Livia Radu, María Menéndez, Henning Müller, Adrian Popescu, and Babak Loni. Div400: a social image retrieval result diversification

dataset. In *Proceedings of the 5th ACM Multimedia Systems Conference*, pages 29–34. ACM, 2014.

[360] Theodora Tsikrika, Jana Kludas, and Adrian Popescu. Building reliable and reusable test collections for image retrieval: The wikipedia task at imageclef. *IEEE MultiMedia*, 19(3):0024, 2012.

[361] Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA, 2008. ACM.

[362] Justus J Randolph. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online Submission*, 2005.

[363] Christof Schuster. A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educational and Psychological Measurement*, 64(2):243–253, 2004.

[364] Alan Mislove, Hema Swetha Koppula, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Growth of the flickr social network. In *Proceedings of the first workshop on Online social networks*, pages 25–30. ACM, 2008.

[365] Meeyoung Cha, Alan Mislove, and Krishna P Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, pages 721–730. ACM, 2009.

[366] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[367] Saeko Nomura, Satoshi Oyama, Tetsuo Hayamizu, and Toru Ishida. Analysis and improvement of hits algorithm for detecting web communities. *Systems and Computers in Japan*, 35(13):32–42, 2004.

[368] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[369] Aixin Sun and Sourav S Bhowmick. Image tag clarity: in search of visual-representative tags for social images. In *Proceedings of the first SIGMM workshop on Social media*, pages 19–26. ACM, 2009.

[370] Xirong Li and al. Learning tag relevance by neighbor voting for social image retrieval. In *ACM MIR 2008*.

[371] Yue Gao and al. Visual-textual joint relevance learning for tag-based social image search. *IEEE TIP*, 22(1), 2013.

[372] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2010.

[373] Duen-Ren Liu, Yu-Hsuan Chen, Wei-Chen Kao, and Hsiu-Wen Wang. Integrating expert profile, reputation and link analysis for expert finding in question-answering websites. *Information Processing & Management*, 49(1):312–329, 2013.

[374] Aditya Pal, Rosta Farzan, Joseph A Konstan, and Robert E Kraut. Early detection of potential experts in question answering communities. In *User Modeling, Adaption and Personalization*, pages 231–242. Springer, 2011.

[375] Reinier H. van Leuken and al. Visual diversification of image search results. In *Proc. of WWW 2009*.

[376] B. Ionescu and al. Div400: A social image retrieval result diversification dataset. *ACM MMSys 2014*.

[377] Adrian Popescu and Nicolas Ballas. Cea list's participation at mediaeval 2012 placing task. In *MediaEval*, 2012.

[378] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[379] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[380] Alessandro Bozzon, Marco Brambilla, Stefano Ceri, Matteo Silvestri, and Giuliano Vesci. Choosing the right crowd: expert finding in social networks. In *Proceedings of the 16th International Conference on Extending Database Technology*, pages 637–648. ACM, 2013.

[381] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(8):1532–1545, 2014.

[382] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46 (5):604–632, September 1999. ISSN 0004-5411. doi: 10.1145/324133.324140. URL http://doi.acm.org/10.1145/324133.324140.

[383] A. Pal and al. Evolution of experts in question answering communities. In *Proc. of AAAI ICWSM'12*, 2012.

[384] L. Mamykina and al. Design lessons from the fastest q&a site in the west. In *Proc. of CHI'11*, pages 2857–2866, 2011. ISBN 978-1-4503-0228-9.

[385] Brian. D. Davison, Apostolos Gerasoulis, Konstantinos Kleisouris, Yingfang Lu, Hyun ju Seo, Wei Wang, , and Baohua Wu. Discoweb: Applying link analysis to web search. In *Proceedings of the 8th International Conference on World Wide Web*, pages 148–149, 1999.

[386] Jun Zhang, Mark S. Ackerman, and Lada Adamic. Expertise networks in online communities: Structure and algorithms. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 221–230, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-654-7. doi: 10.1145/1242572.1242603. URL http://doi.acm.org/10.1145/1242572.1242603.

[387] B. Hanrahan and al. Modeling problem difficulty and expertise in stackoverflow. In *Proc. of ACM CSCW'12*, pages 91–94, 2012.

[388] A. Pal and al. Early detection of potential experts in question answering communities. In *Proc. of UMAP'11*, pages 231–242, 2011.

[389] Aditya Pal and Scott Counts. What's in a @name? how name value biases judgment of microblog authors. In *Proc. of ICWSM'11*, 2011.

[390] T. Deselaers and al. Features for image retrieval: an experimental comparison. *Inf. Retr.*, 11(2):77–107, April 2008.

[391] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=944919.944937.

[392] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proc of IJCAI'07*, pages 1606–1611, 2007.

[393] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4): 77–84, 2012.

[394] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit (consulted on 11/02/2013) http://mallet.cs.umass.edu. 2002.

[395] Thorsten Joachims. Training linear svms in linear time. In *Proc. of ACM KDD'06*, pages 217–226. ACM, 2006.

[396] V.I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, 1966.

Résumé

Le travail présenté dans cette Thèse est placé au carrefour entre l'utilisation de données de Web dans la fouille d'images et la crédibilité des sources dans les plates-formes de partage d'images. Il vise à apporter des découvertes importantes aux deux domaines et fournir un lien prometteur entre deux secteurs séparés de recherche. Les cadres théoriques et les résultats expérimentaux que nous exposons en détail peuvent servir aux chercheurs avec des intérêts différents: i) ceux venant de la communauté de la fouille multimédia, en introduisant des représentations d'images sémantiques efficaces construites à partir des ressources d'images librement disponibles et ii) les chercheurs intéressés à la qualité de données de Web et à la crédibilité des sources, en proposant une étude de crédibilité dans le domaine multimédia et en évaluant des applications pratiques d'estimations de crédibilité d'utilisateur.

Nous proposons un cadre de classification d'images évolutif, qui exploite des classifieurs linéaires binaires. Pour implémenter ce framework, nous comparons deux sources de données: un grand jeu de données d'images manuellement annotées (c.-à-d. ImageNet) et les groupes de Flickr. Comme la deuxième ressource est recueillie des images de Web, une partie méthodologique indispensable de travail détaille des méthodes qui réduisent le bruit inhérent à la collection. Dans une section expérimentale prolongée, nous montrons que les descripteurs sémantiques proposés non seulement améliorent la performance de recherche sur trois collections d'images bien connues, quand ils sont comparés à l'état d'art, mais aussi offrent une amélioration significative des temps de recherche.
Ensuite, nous définissons le concept de crédibilité des utilisateurs et l'appliquons aux utilisateurs de Flickr. Nous proposons 66 traits qui peuvent servir comme estimateurs pour la crédibilité des utilisateurs. Nous introduisons des traits basés sur le contexte, aussi que des traits basés sur le contenu extrait des différentes données de Flickr. Nous évaluons les traits proposés tant sur une collection publiquement disponible que sur un nouveau jeu de données, que nous introduisons dans cette Thèse. Finalement, nous montrons l'utilité des estimateurs de crédibilité dans deux scénarios d'application: par les introduisant dans un pipeline de diversification d'un système de recherche d'images et le fait de les utiliser comme des traits dans des modèles d'apprentissage pour la classification d'expertise et des tâches de recherche d'experts.

Ce travail contribue à une meilleure compréhension et à un modelage d'intelligence sociale pour les tâches de traitement de l'information. Nous nous sommes concentrés sur la recherche d'images et l'estimation de crédibilité dans des plate-formes multimédia, mais les méthodes proposées sont aussi pertinentes pour d'autres applications.

Mots-clés : Crédibilité d'utilisateurs, Traitement des données du Web, Représentation sémantique de l'image, Recherche d'images par le contenu, Diversification visuelle des résultats, Concepts visuels

Abstract

While research in visual and multimedia recognition and retrieval has significantly benefited from manually labeled datasets, the availability of such resources remains a serious issue. Manual annotation is still a cumbersome task, especially when it is conducted on large datasets. A promising way to circumvent the lack of annotated data is to use images shared on multimedia social networks, such as Flickr. One of the main drawbacks of user-contributed collections is that a part of images annotations is not directly related to the visual content, rendering them less useful for image mining.
The work presented in this Thesis is placed at the crossroads between the use of Web data in image mining and source credibility in image sharing platforms. It aims at bringing novel findings to both domains and furnishing a promising link between two separate fields of research. The theoretical frameworks and experimental results we detail can benefit both i) researchers coming from the multimedia mining community, by introducing efficient semantic image representations built from freely available image resources and ii) researchers interested in Web data quality and source credibility, by proposing a study of credibility in the multimedia domain and testing practical applications of user credibility estimates.

We propose a scalable image classification framework that exploits binary linear classifiers. To implement this framework, we compare two data sources: a large manually annotated image dataset (i.e. ImageNet) and Flickr groups. For the second, we details methods that reduce the noise inherent to a Web collection. In an extended experimental section, we show that the proposed semantic features not only improve the retrieval performance on three well known image collections, when compared to state of the art image descriptors, but also offer a significant improvement of retrieval time.

We then define the concept of user tagging credibility and apply it to Flickr users. We propose 66 features that can serve as estimators for user credibility. We introduce both context and content based features extracted from various Flickr data. We evaluate the proposed features both on a publicly available dataset and new dataset, which we introduce in this Thesis. Finally, we showcase the use of credibility estimates in two application scenarios: embedding them in an image diversification pipeline and using them as features in machine learning models for expertise classification and expert retrieval tasks.

This work contributes to a better understanding and modeling of social intelligence for information processing tasks. We focused on image retrieval and multimedia credibility estimation but the methods proposed here are also relevant for other applications, such as image annotation and Web data quality control.

Keywords : User credibility, Expert retrieval, Web data processing, Semantic image representation, Visual concept classification, Content based image retrieval, Image retrieval diversification

TELECOM Bretagne

n° d'ordre : 2015telb0387
Télécom Bretagne
Technopôle Brest-Iroise - CS 83818 - 29238 Brest Cedex 3
Tél : + 33(0) 29 00 11 11 - Fax : + 33(0) 29 00 10 00