



Text-Based Ephemeral Clustering for Web Image Retrieval on Mobile Devices (version 1)

José G. Moreno

► To cite this version:

José G. Moreno. Text-Based Ephemeral Clustering for Web Image Retrieval on Mobile Devices (version 1) . Computation and Language [cs.CL]. Université de Caen Basse-Normandie, 2014. English. NNT: . tel-01102604

HAL Id: tel-01102604

<https://hal.science/tel-01102604>

Submitted on 13 Jan 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE CAEN BASSE NORMANDIE

U.F.R. de Sciences

ÉCOLE DOCTORALE SIMEM

T H È S E

Présentée par

M. Jose Guillermo MORENO FRANCO

soutenue le

10 Décembre 2014

en vue de l'obtention du

DOCTORAT de l'UNIVERSITÉ de CAEN

Spécialité : Informatique et applications

Arrêté du 07 août 2006

Titre :

**Text-Based Ephemeral Clustering for
Web Image Retrieval on Mobile Devices.**

Laboratoire : Groupe de recherche en informatique, image, automatique et
instrumentation de Caen (GREYC)

Jury

M. Massih-Reza AMINI	Professeur des Universités	Université Joseph Fourier	<i>Rapporteur</i>
M. Adam JATOWT	Associate Professor	Kyoto University	<i>Rapporteur</i>
Mme. Béatrice DAILLE	Professeur des Universités	Université de Nantes	<i>Examineur</i>
M. Olivier FERRET	Chercheur	Institut CEA LIST	<i>Examineur</i>
M. Guillaume CLEUZIOU	Maître de Conférences	Université d'Orléans	<i>Examineur</i>
M. Marc SPANIOL	Professeur des Universités	Université de Caen	<i>Examineur</i>
M. Gaël DIAS	Professeur des Universités	Université de Caen	<i>Directeur de thèse</i>

Abstract

In this thesis, we present a study about Web image results visualization on mobile devices. Our main findings were inspired by the recent advances in two main research areas - Information Retrieval and Natural Language Processing. In the former, we considered different topics such as search results clustering, Web mobile interfaces, query intent mining, to name but a few. In the latter, we were more focused in collocation measures, high order similarity metrics, etc. Particularly in order to validate our hypothesis, we performed a great deal of different experiments with task specific datasets. Many characteristics are evaluated in the proposed solutions. First, the clustering quality in which classical and recent evaluation metrics are considered. Secondly, the labeling quality of each cluster is evaluated to make sure that all possible query intents are covered. Thirdly and finally, we evaluate the user's effort in exploring the images in a gallery-based interface. An entire chapter is dedicated to each of these three aspects in which the datasets - some of them built to evaluate specific characteristics - are presented.

For the final results, we can take into account two developed algorithms, two datasets and a SRC evaluation tool. From the algorithms, Dual C -means is our main product. It can be seen as a generalization of our previously developed algorithm, the AGK -means. Both are based in text-based similarity metrics. A new dataset for a complete evaluation of SRC algorithms is developed and presented. Similarly, a new Web image dataset is developed and used together with a new metric to measure the users effort when a set of Web images is explored. Finally, we developed an evaluation tool for the SRC problem, in which we have implemented several classical and recent SRC metrics.

Our conclusions are drawn considering the numerous factors that were discussed in this thesis. However, additional studies could be motivated based in our findings. Some of them are discussed in the end of this study and preliminary analysis suggest that they are directions that have potential.

Résumé

Dans cette thèse, nous présentons une étude sur la visualisation des résultats Web d'images sur les dispositifs nomades. Nos principales conclusions ont été inspirées par les avancées récentes dans deux principaux domaines de recherche – la recherche d'information et le traitement automatique du langage naturel. Tout d'abord, nous avons examiné différents sujets tels que le regroupement des résultats Web, les interfaces mobiles, la fouille des intentions sur une requête, pour n'en nommer que quelques-uns. Ensuite, nous nous sommes concentré sur les mesures d'association lexicale, les métriques de similarité d'ordre élevé, etc. Notamment afin de valider notre hypothèse, nous avons réalisé différentes expériences avec des jeux de données spécifiques de la tâche. De nombreuses caractéristiques sont évaluées dans les solutions proposées. Premièrement, la qualité de regroupement en utilisant à la fois des métriques d'évaluation classiques, mais aussi des métriques plus récentes. Deuxièmement, la qualité de l'étiquetage de chaque groupe de documents est évaluée pour s'assurer au maximum que toutes les intentions des requêtes sont couvertes. Finalement, nous évaluons l'effort de l'utilisateur à explorer les images dans une interface basée sur l'utilisation des galeries présentées sur des dispositifs nomades. Un chapitre entier est consacré à chacun de ces trois aspects dans lesquels les jeux de données - certains d'entre eux construits pour évaluer des caractéristiques spécifiques - sont présentés.

Comme résultats de cette thèse, nous sommes développés : deux algorithmes adaptés aux caractéristiques du problème, deux jeux de données pour les tâches respectives et un outil d'évaluation pour le regroupement des résultats d'une requête (SRC pour les sigles en anglais) . Concernant les algorithmes, Dual C -means est notre principal contribution. Il peut être vu comme une généralisation de notre algorithme développé précédemment, l' AGK -means. Les deux sont basés sur des mesures d'association lexicale à partir des résultats Web. Un nouveau jeu de données pour l'évaluation complète d'algorithmes SRC est élaboré et présenté. De même, un nouvel ensemble de données sur les images Web est développé et utilisé avec une nouvelle métrique à fin d'évaluer l'effort fait pour les utilisateurs lors qu'ils explorent un ensemble d'images. Enfin, nous avons développé un outil d'évaluation pour le problème SRC, dans lequel nous avons mis en place plusieurs mesures classiques et récentes utilisées en SRC.

Nos conclusions sont tirées compte tenu des nombreux facteurs qui ont été discutés dans cette thèse. Cependant, motivés par nos conclusions, des études supplémentaires pourraient être développés. Celles-ci sont discutées à la fin de ce manuscrit et notre résultats préliminaires suggère que l'association de plusieurs sources d'information améliore déjà la qualité du regroupement.

Contents

Abstract	1
Abstract	3
1 Introduction	7
1.1 Research Objective	10
1.2 Problem Definition	11
1.3 Thesis motivation	12
1.4 Proposed Solution	12
1.5 Limitations	13
1.5.1 Exploitable Content	13
1.5.2 Repeatability	13
1.5.3 Ephemeral Clustering	13
1.6 Thesis organization	14
1.7 Conclusions	15
2 Related Work and Preliminary Investigations	17
2.1 Introduction	17
2.2 Web Image Retrieval on Mobile Devices	18
2.2.1 Text-based Web Image Clustered Visualization	21
2.2.2 Text-based vs Query Log Clustering by Expansion	22
2.2.3 Evaluating Clustering by Expansion	25
2.2.4 Results and Discussion	28
2.3 Ephemeral Clustering and their baselines	30
2.3.1 Search Results Clustering and Related Work	31
2.3.2 Evaluation Metrics and Datasets	34
2.3.3 Current Results	37
2.3.4 Understanding Baseline Algorithms and Metrics	38
2.4 Conclusions	42
3 Improving Text-based Ephemeral Clustering	45
3.1 Introduction	45
3.2 Polythetic Search Results Clustering	47
3.2.1 Intuitive Idea	47
3.2.2 Polythetic Clustering	47
3.2.3 Collocation Measures	50
3.3 Stopping Criterion	50
3.4 Text-based Evaluation	52
3.4.1 Text Processing	53

3.4.2	Intrinsic Evaluation	54
3.4.3	Comparative Evaluation	54
3.5	Conclusions	57
4	Including External Information in Ephemeral Clustering	59
4.1	Introduction	59
4.2	Dual C-means Algorithm	61
4.2.1	General Model	61
4.2.2	Instantiation in the SRC Context	62
4.3	The WEBSRC401 Dataset	65
4.4	Clustering Evaluation	66
4.4.1	Evaluation of SRC	66
4.4.2	Experimental Setups	68
4.4.3	Clustering Results	70
4.5	Labeling Evaluation	73
4.6	Conclusions	76
5	Web Image Clustered Visualization in a Mobile Context	77
5.1	Introduction	77
5.2	Text-Based Web Image Search Results Clustering	79
5.2.1	A Gallery-Based Interface	79
5.3	The Web Image SRC Dataset	80
5.4	Evaluation and Results	81
5.4.1	Clustering Performance	81
5.4.2	Wasted Space-Interface	82
5.5	Conclusions	85
6	Conclusion and Future Work	87
6.1	Conclusions and Contributions	87
6.2	Preliminary studies in Future Directions	91
6.2.1	A Novel Non-Content Based Direction	91
6.2.2	A Multi-objective Search Results Clustering Direction	92
6.3	Final Remarks	93
	List of figures	95
	List of tables	97
	Bibliography	99

Chapter 1

Introduction

Communication is one of the main actions in our daily lives. It consists of transmitting our ideas, thoughts and/or knowledge to others. In fact, our actual knowledge about the historical events or previous discoveries comes in the form of one of the most used communication channels, in writing. In past years, collections of written documents were grouped together in specific places which we now refer to libraries. Actually, libraries are considered as invaluable centers of knowledge due to the huge amount of information available in these places. Accessing all of this information is a major interest in many aspects. Fortunately, because of the apparition of the Web and when digital versions are available, accessing these documents has become much easier, mainly because content can be available without geographical or temporal restrictions. Electronic version of the libraries, the eLibraries, is a cheap and alternative option to facilitate book access. Certainly, we frequently ignore where the desired book could be located, but with an eLibrary we know that a Web connexion suffices to access it.

Indeed, when comparing with the physical exploration of information (e.g. going through the pages of a book) the access of information through the Web is each time more and more affordable even if the accessed content is exactly the same. So, what is more frequently changing is the way that the content is accessed, but not the content by itself. Consider the Google Books¹ Web site, in it we can search for a desired piece of information contained in a book by only writing down what we have in mind instead of manually exploring the book until discovering the right section.

However, not always the desired information can be easily described in an accurate query or many results could be considered as relevant. To address this situation, the use of manual categories is a recurrent solution. Nowadays, many repositories permit the exploration of information by manually constructed categories which enhance the exploration task of the user. For example in Figure 1.1, we can see the results of products sold on

¹<http://books.google.com> [Last access: 20/06/2014].

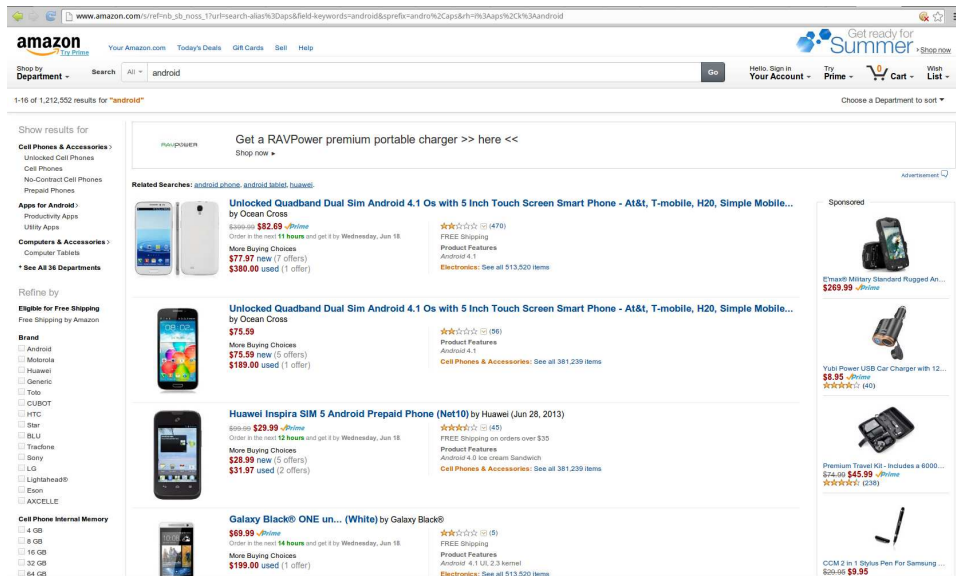


Figure 1.1: Search of the query “android” in amazon.com [Last access: 17/06/2014].

a commercial Web site² obtained by using the query word “android”. In this case, the Web site provides us a list of products relevant to the request. Note that only the top four relevant products are shown in the first screen. However, on the left side a set of options is listed. They are presented under the label “Refine by” and it corresponds to information previously registered about the products’ specifications. For example, if we select the option “generic” in the “brand” section the list of results is updated with only the products labeled as “generic”. This interface facilitates the previous task of the user after typing the query - helping to accurately redefine the search. However, it is only possible in specific cases when the resulting items include a manually supplied classification.

In fact, even when manual categories are a useful way to store information, but it is often not suitable on the Web scale. Two main factors must be considered. First, the high rate of change present in the Web which demands daily updates. And second, the dependency on the context of the word meanings which can generate misunderstandings when a user explores on a large set of categories. Indeed, in a typical search engine, Web pages are collected through Web crawling and Indexing - the two main phases in charge of exploring and storing Web content. These phases are often performed automatically and can not be manually processed given the always increasing amount of Web documents. For this reason, categorical structures manually updated of the Web are not common³. The second reason is the high cost of manually updating. Nevertheless, applying adapted categorical structures or on-the-fly categories built through the use of clustering techniques is a broadly accepted strategy to automatically categorize Web documents. Indeed, Web document clustering is a well-known strategy and has been explored in different commercial systems

²http://www.amazon.com [Last access: 20/06/2014].

³DMOZ is a manually categorized version of the Web content. However, as any other manual built resource it is often out-to-date.

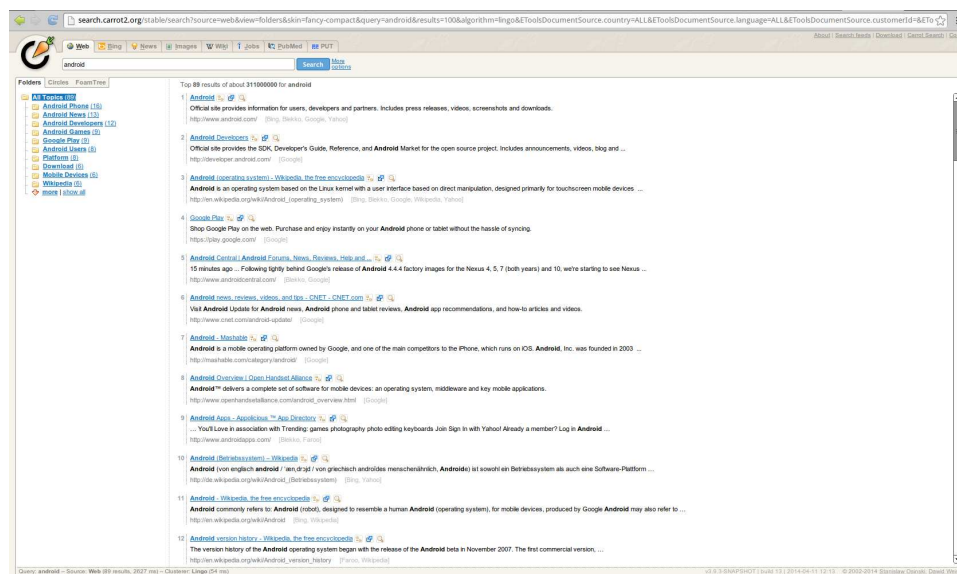


Figure 1.2: Search of the query “android” in carrot2.org [Last access: 17/06/2014].

such as: Yippy⁴, Clusty⁵ and Carrot2⁶. In which, instead of presenting a list of results for a given query, the system presents the results grouped in automatically constructed categories. In Figure 1.2, we see the clustered results obtained for the same previous query using Carrot2. In this case, similar to Figure 1.1, at the left we have a list of categories of Web results. However, these categories correspond to the automatically extracted labels of the cluster groups of Web documents.

Search engines and Web access are too much integrated into our lives to the point that we can consider their use as part of our daily routines. Based on recent information⁷, we can estimate that 77% of the population in the developed world are users of the Web. Surely, the main way people access the Web is through Search Engines (SE). Their use is so well integrated into our concept of Web searching that we frequently use the expression “google⁸ something” (or “google it!”) as the action of performing a search with a SE. It could certainly be hard for infrequent users to properly differentiate the Web content from a SE because they confuse it with the Web content. This misunderstanding is clear when someone say that found something in google when in reality the content is in another Web site but it was found using google.

In parallel, mobile devices like smartphones are replacing desktop computers as popular devices to access the Web. As a consequence, the use of mobile devices is also replacing other devices to perform searches. Estimations indicate that by 2015 more users will access the Internet through mobile devices than desktops computers which supports the

⁴<http://www.yippy.com> [Last access: 20/06/2014].

⁵<http://www.clusty.com> [Last access: 20/06/2014].

⁶<http://www.carrot2.org> [Last access: 20/06/2014].

⁷Internet Usage http://en.wikipedia.org/wiki/Global_Internet_usage [Last access: 13/06/2014].

⁸GoogleTM is one of the mayor commercial SE available in the Web.

idea that searching on mobile devices will be a major concern in the developing of new capabilities for search engines. Consequently, we can expect that in the near future the primary way to access the Internet will be through mobile devices. So, new ways to use and interact with Web results is still attracting some attention.

However, Web search itself is already a difficult task. First, query definition is a subjective user task. Two users can use the same query string to refer to completely different concepts or intentions. Some tools have been implemented by the search engines to help with this task. The two more frequent options are query suggestions and query corrections. In both cases, the search engines propose some alternative strings related to a given query or partial query. This completion of the query is helpful to the users, however it can also mislead the users in some ways to imprecisely select a query. On the other hand, some difficulties when a search is performed could be caused by the search engine. One of the most frequent ones is because of their interfaces. Nowadays, search engines offer similar search interfaces even when different kinds of content are searched or different devices are used. Indeed, searching for a Web image or a Web site in commercial search engines differentiates slightly in the interface. Web sites are presented in a top-to-bottom list and Web images are presented in a top-to-bottom grid. Indeed, for many operating systems the desktop and mobile interfaces are almost the same. In certain situations, the Web interface of a SE is degraded depending of the used operating system⁹ to access the SE. This unfortunate situation has recently been inspiring many works focused on exploring alternative solutions for Web search interfaces on mobile devices.

Following these interesting topics, this thesis presents a research effort in understanding and developing of novel text-based clustering techniques for the exploring of Web image search results with mobile devices. The topic is ambitious but to ground our proposal we present our research objectives, the problem definition, the motivations and the main limitations of the thesis in the following sections.

1.1 Research Objective

The aim of this thesis is to explore and propose novel techniques to combine *word frequencies* information and *clustering algorithms* into a unique framework to enhance the user experience when a set of Web image results is explored on a *mobile device*. As this is not an unknown area, we relied on previous studies which showed that clustering enhance the Web results exploration. However, we concentrate our efforts in the analysis of labels construction and cluster quality along with their impact in user experience on mobile devices.

⁹For example, when using Google Images from a Windows mobile phone [Last access: 20/06/2014].

1.2 Problem Definition

Our general problem consists of grouping together sets of documents. However, not only clustered documents must be provided. The main problem is the assignation of a label to each of the obtained clusters. Additionally, two main restrictions must be considered. First, Web image search results are the target documents in our research. Moreover, only text-content could be used as available information. Namely, we only use the textual information associated to Web images (e.g. snippets) and discard any visual information. Second, results must be displayed on mobile devices such as smartphones. For this reason, classical interactions available in smartphones are considered as exploitable characteristics of the problem.

Research questions

In the context of mobile search results exploration, the establishing of an interface is an interesting challenge. However, some studies have been proposed in this direction. Supposing we have a fixed interface in which a set of image search results are explored, we are interested in addressing the following two research questions:

- Q1. Is it better to use automatic labeling strategies or query logs to describe clusters of images?
- Q2. How much effort is need to explore a set of Web images in a mobile device?

Repeatable research is one of the hardest challenges in computer science. Indeed, reproducing an experiment includes many phases: coding of state-of-the-art algorithms, using standard broadly used dataset collection, and finally, evaluating performances by using adapted evaluation metrics. However, some of the unexplored or unknown parameters can deal with completely different results when compared with original works. Following that, we propose to cope with the following research questions.

- Q3. Are known algorithms correctly evaluated in terms of metrics and configurations?
- Q4. Is it possible to measure labeling quality and clustering quality in a reproducible experiment?

And finally, inspired by works on word context vectors, we expect to revisit some of the classical problems in search results clustering: labeling and cluster size definition.

- Q5. Can query logs be integrated into ephemeral clustering techniques?
- Q6. Is it possible to automatically define the numbers of underlying topics in a query?

1.3 Thesis motivation

This thesis is mainly inspired by the advances in areas such as Word Frequency Analysis (Collocation Measures), Ephemeral Clustering and Mobile Internet Access. But a study that combines each of these areas strengths in a unique framework is still lacking.

Word Frequencies Analysis The use of collocation metrics¹⁰ permits the analysis of non-contiguous words that share certain properties about their document frequencies. As a recurrent idea, our work is considering that an individual word could be automatically associated with other words through the use of collocation techniques. This is an alternative technique to the classical use of string matching.

Ephemeral Clustering In previous studies, it has been proven that clustering of documents obtained from SEs improves user satisfaction when interacts with Web results. Our efforts will be concentrated in improving the knowledge about these techniques, mainly focusing on the comprehension of evaluation techniques for cluster quality and label quality.

Mobile Web Image Retrieval The growing use of mobile devices and tools to access Web image content is modifying the way that users interact with Web results. The application of ephemeral clustering techniques over mobile interfaces is applied in this study.

1.4 Proposed Solution

Our research produced a set of scientific contributions such as ephemeral clustering algorithms, evaluation metrics, research datasets and tools available for the research community. These contributions are briefly presented in this section.

Clustering algorithms Two main algorithms are presented in this theses, the *AGK*-means algorithm presented in the Chapter 3 and the Dual *C*-means algorithm presented in Chapter 4.

Evaluation metrics One evaluation metric is proposed in Chapter 5 to measure the users' effort when a set of images is explored on a mobile device. Moreover, other metrics typically used in intent mining are proposed for the labeling quality evaluation.

¹⁰We use the term collocation metric to refer to association measures used for the collocation extraction techniques.

Research datasets and tools Two ephemeral clustering datasets were built and they are presented in Chapters 4 and 5. The first one text ephemeral clustering dataset and the second one is a Web image ephemeral clustering dataset. Both are publicly available. Finally, we implement an evaluation tool presented in Chapter 2 that allows the calculation of several ephemeral clustering metrics.

1.5 Limitations

In order to clarify the context of this study, we present two main limitations considered during its development.

1.5.1 Exploitable Content

The task of clustering Web images includes many exploitable sources such as text content, image content, metadata information or link information of the Web site that contains the image. However, we chose text content as our main target. As similar studies in ephemeral clustering, Web snippets are our preferred text content information. For that reason, our explored techniques are based mainly on text content. However in Section 6.2, we discuss the possible use of link information and image content in future work.

1.5.2 Repeatability

Independently of the studied area, experiments conducted with users are hard to accurately reproduce. Many factors could induce the users to different outputs. To address this situation, many experiments to measure the performance of a set of algorithms are based on previously annotated datasets. In this thesis, we aim to present repeatable experiments using standard metrics and datasets. Although in Chapter 2 a small experiment with users is presented, the rest of this thesis is focused on reproducible experiments. In particular, we propose that some users' interaction experience are simplified in a model in order to use reproducible metrics. However, comparison to real situations is expected. This is a common situation in this kind of works where both "reproducible experiments" and "user evaluation" are needed to confirm all research advances.

1.5.3 Ephemeral Clustering

Some of the limitations are not only present in this thesis, but also in the original kind of techniques. Ephemeral clustering uses the Web results as input to generate a group of documents. This is often remarked as an advantage because it allows a quick adaptability to new documents. However, it is also considered as a drawback because the clustering

process must be performed each time that new Web results list is obtained and previous clustered results are discarded. Although this limitation, the adaptability of the ephemeral clustering techniques remains as an interesting property to take advantage.

1.6 Thesis organization

Chapter 2 This chapter presents the main ideas of our research. Firstly, the recent shift in human-computer interaction from desktop to mobile devices and the actual needs of new interfaces for Web image search results exploration. Secondly, an exhaustive evaluation of the existing research in text-based clustering. In particular, our proposed configuration - the cascade configuration - deals good results, which proves comparable results to recent text-based state-of-the-art algorithms.

Chapter 3 This chapter shows our advances in Ephemeral clustering - the task of clustering Web search results. Within this context, we propose a new methodology based on collocation measures. Consequently, documents are not represented in a classical space vector model instead documents are represented by more relevant constituents of them. Obtained results show that the proposed criterion outperforms all reported text-based approaches and that is a suitable algorithm for the search results clustering problem.

Chapter 4 This chapter presents our proposal to use external resources in order to improve labeling quality. To do that, we developed a new algorithm called *Dual C-Means*, which provides a theoretical background for clustering in different representation spaces. Its originality relies on the fact that external resources can drive the clustering process as well as the labeling task in a single step. Results demonstrates its significant advantages over traditional clustering and labeling techniques.

Chapter 5 In this chapter, we present the developed algorithm - in the case of Web image results clustering - when a mobile device is used to explore the results. In order to leverage users' efforts, we used our developed ephemeral clustering algorithms and we evaluate them using a new metric to evaluate the mismatch of the used space-interface between the ground truth and the cluster distribution obtained by the ephemeral clustering algorithms. Results show that our solutions are the best option when all evaluated factors are considered.

Chapter 6 In this chapter, we summarize the content of the thesis and highlight the main contributions obtained through the development of the work. Indeed, research achievements are discussed as well as newly built datasets and their impact in future

research. Finally, we discuss some possible directions for further work with preliminary experiments.

1.7 Conclusions

In this first chapter, we presented the main objective of this research and defined our contributions for the research community. Similarly, main limitations are presented and explained. A brief description for each of the following chapters is also presented.

In next chapter, we present the related work, state-of-the-art results and our preliminary advances in ephemeral clustering.

Chapter 2

Related Work and Preliminary Investigations

Contents

1.1 Research Objective	10
1.2 Problem Definition	11
1.3 Thesis motivation	12
1.4 Proposed Solution	12
1.5 Limitations	13
1.5.1 Exploitable Content	13
1.5.2 Repeatability	13
1.5.3 Ephemeral Clustering	13
1.6 Thesis organization	14
1.7 Conclusions	15

2.1 Introduction

These days, searching on the Web is becoming a daily activity since Web search is the most used way to access Web content. Currently, a commercial Search Engine is processing around three billion searches every day including a broad number of platforms. However, more commonly used Search Engines present their results as a list of relevant Web results. To address this problem, many studies have been proposed for Web page search results clustering [Zamir and Etzioni, 1998, Carpineto et al., 2009] and some for Web image search results clustering [Cai et al., 2004, Ding et al., 2008].

Ephemeral clustering, also known as search results clustering or post-retrieval clustering (PRC), has been broadly used as a way to improve the search results exploration task.

SRC consists in the clustering Web search results. We dedicate this chapter to the drawing of the most well-known state-of-the-art techniques for text-based search results clustering in both following contexts: mobile device applications for Web image results and pure text algorithms. Our idea is that text-based techniques are suitable for the mobile device context by the use of query expansion. However, query expansion techniques modify the obtained results thus affecting the original result list obtained from the search engine. For this, we also present a set of purely text-based algorithms for the Web Search Results Clustering task. In this context, we analyze their performances over different configurations of well-known baseline techniques. Note that the scope of this work is oriented towards text-based techniques, for that, many works that use visual-based techniques are excluded or only briefly mentioned (see Section 1.5.).

In Section 2.2, we present our preliminary work in mobile device visualization of Web image results and a clustering by expansion technique that allows us to draw the interest of clustering interfaces in the mobile devices context. Note that this first set of experiments is not a traditional search results clustering algorithms, but they allow us to motivate their use in the context of Web image results exploration. Section 2.3 is completely dedicated to traditional search results clustering algorithms. In particular, we carry a set of preliminary experiments to better understand the capabilities and performances of well-known baselines for text-based search results clustering.

2.2 Web Image Retrieval on Mobile Devices

In recent years, the growing number of mobile devices with Internet access has changed how people access Web content as well as their interaction with such content [Kamvar and Baluja, 2006]. In particular, new applications are developed to support native features, which take into account device peculiarities to enhance user interaction. However, performing Web search in a common commercial search engine is still made in a similar way as in desktop computers, i.e. a simple list of ranked results is shown to the user. But, ranked lists are not suitable for exploration and selection of relevant results, especially on mobile devices where small screen size is a tendency.

In the context of mobile Web image search, enhanced user interaction is a crucial task as mobile devices have small screens that restrict the number, quality and size of Web images results displayed. Moreover, on mobile devices browsing must be as simple as possible due to interaction limitations of such devices. For example, when a search is performed and a simple ranked list of Web image results is displayed, the user must browse the entire list to look for the expected image. Usually, the longer this exploratory phase is, the higher the users networking consumption grows. In Figure 2.1, we show the Android interface for Google Web image search results when the query “colombia” is performed. Note that the Web image results are presented in form of grid and in the bottom a numeric



Figure 2.1: Web image results for the query “colombia” in the Android platform.

list of links are included to allow the page list navigation. This example clearly shows the difficulties when a Web image search is performed. To address these problems, image collection visualization techniques have been used within novel image retrieval systems to allow new user interactions [Cai et al., 2004, Ding et al., 2008]. However, most of these systems are implemented for desktop computers¹ and do not take into account restrictions present when the user uses a mobile device.

Recently, the image search engine service provided by Google, Google Images, updated the way that mobile users² may interact with the application [Google, 2010]. But the new interaction still has problems such as screen space waste and page jumping difficulties to mention just but a few.

In the context of desktop computers, researchers have proposed new user interaction solutions using artificial intelligence techniques to enhance Web image retrieval such as the “Sort by subject” facility proposed in [Google, 2011b]. However, previous works on human-computer interaction have shown that mobile user needs are different than when they work in desktop computers [Sohn et al., 2008]. Indeed, these differences can be significantly relevant depending of the mobile devices which is used [Kamvar et al., 2009].

¹With some recent exceptions [Tolchinsky et al., 2012].

²Iphone and Android users in particular.

Although the desktop solutions could be easily implemented in mobile devices interfaces, actual methodologies show specific drawbacks for that option such as language dependency [Google, 2011b], fixed image collections [Chatzichristofis et al., 2010], or predefined categorization that restricts their use in a Web search scenario.

Performing a successful text query is still a challenge from both, a purely research point of view or a commercial purpose. To enhance the search process, search engines provide a common tool in Web search retrieval called query suggestion. Query suggestions have been used in a variety of ways to help the user query definition process. There are two types of query suggestions facilities: (1) pre-submission suggestions that appear as an auto-complete box and (2) post-submission suggestions, which are usually links used to redirect to new queries. Usually, post-submission suggestions are shown by search engines after a text like *“Including results for ...”*, *“Showing results for ...”* or *“We have included ...”* and some others offer more than one result shown as *“Searches related to ...”*. In particular, typography errors can be solved with post-submission suggestions and more complex problems related to (1) the query such as semantic disambiguation, (2) the device such as typography difficulties or (3) the user such as users’ time consumption are addressed by search engines using pre-submission suggestions³. In particular, typography difficulties are common in mobile devices because the keys in these keyboards are used for different letters or because of their small size [Kamvar and Baluja, 2007]. These pre- and post-submission suggestions are already implemented in mobile devices, however picking the right pre-submission suggestion is not a easy task [Paek et al., 2009].

Particularly, in the commercial context, Google suggestions are extracted from query logs captured from previous users’ search activities [Google, 2011a]. In the same way, Yahoo! suggestions come from users’ query definition behaviors [Yahoo, 2011]. In the mobile context, the use of query logs as suggestions can reduce keystrokes in common searches and help to reduce user time consumption for query determination [Paek et al., 2009] as well as it can increase the number of characters used in the query to improve the precision of the search [Kamvar and Baluja, 2008]. Given the use of query logs information, these approaches are highly dependent on previous users’ search activities and in fact, are unusable in the context of unknown queries.

As a consequence, to explore the possibilities in Web image search results organization when mobile devices are used, we propose two customized mobile image visualizations based on clustering by query expansion: (1) through an ephemeral clustering technique as presented in [Carpineto and Romano, 2009] and (2) through the exploration of query logs to generate query suggestions as cluster names. A preliminary experiment is performed with frequently used text queries to evaluate the results under a user based evaluation. Moreover, we also propose an automatic evaluation of the same experiment.

³This is possible when a user has previously used the same device. The suggestion is usually presented in a different color, i.e. in purple.

2.2.1 Text-based Web Image Clustered Visualization

The most popular way to access to the image content in the Web is using text-based queries [Datta et al., 2008]. Normally, popular image search engines such as Yahoo! and Google use the text content from the Web page where the image was found for query matching. Usually, information such as the file name, caption or surrounding text gives more chances to achieve the top of the returned results for text-based searches. Even when services like TinEye⁴, GazoPa⁵ or Google Images⁶ offer content-based visual searches, the keyword search is still the preferred way. This can be due to the high accuracy obtained by text based systems when compared to with visual-based systems [Müller et al., 2010].

Web cluster retrieval has been an interesting way to visualize search results and has been investigated by researchers with different approaches from text retrieval techniques [Dias et al., 2011] [Zeng et al., 2004] to image retrieval methodologies [Cai et al., 2004] [Wang et al., 2004] [Wang et al., 2007] [Ding et al., 2008], but without notable impact on commercial search engines. However, this feature is actually available in the Google image retrieval interface [Google, 2011b]. Nevertheless, a large number of clusters and imprecise labels are the main reasons to avoid the use of Web cluster retrieval on main search engines [Dias et al., 2011]. Ephemeral Clustering has been used for Web search results classification and has shown good results in previous works for mobile applications. In the context of Web page search, the number of obtained clusters and the cluster label are important factors to expedite the user interaction time expense in the retrieval process.

Text-based and mixed approaches have been used to address the image Web cluster retrieval problem. Mixed approaches include visual feature extraction and analysis as well as text features. One of the most interesting works tackling the mixed approach is proposed by [Ding et al., 2008]. They first reformulate the text query with frequent co-occurring key phrases. Then, they organize the images inside text clusters obtained through ephemeral clustering by visual contents. Although, the visual similarity between the images can be guaranteed, semantically related images may not belong to the same visual cluster and many clusters may be generated to include uncategorized images. Moreover, [Ding et al., 2008] use the text ephemeral clustering algorithm proposed by [Zeng et al., 2004], which over-generates clusters and builds an *Other Topics* cluster for unsolved web pages. As such, many possible results are lost as the cluster can not be used reliably to show semantically related images. Comparatively, [Wang et al., 2004] address the visual semantic gap including visual phrasal analysis but do not provide names to the clusters as they do not use ephemeral clustering but just combine text and visual in-

⁴<https://www.tineye.com> [Last access: 19/06/2014].

⁵Recently shutdown <http://www.gazopa.com/> [Last access: 19/06/2014].

⁶<http://images.google.com/> [Last access: 19/06/2014].

formation in terms of vectors for classical centroid clustering. In [Cai et al., 2004], the authors propose a novel technique to combine Web page structure, text information and low level image features. The user time consumption problem is addressed using a cluster presentation. In particular, they produce three different image representations: a textual feature, a visual feature and a link-graph feature. Their algorithm consists in an initial text cluster where a spectral technique is applied to define the appropriate number of clusters. Then, the Web structure and visual information are combined, which show interesting results for one query. The major drawback of this work is the fact that they do not propose a general evaluation with more than one query. As such, the importance of the approach can not be accessed.

In parallel, [Wang et al., 2007] propose a Web image search strategy using text clustering of web search results to reformulate the queries. This approach is based on text analysis only as image textual indexing is supposed to be processed in a first step. A set of new related queries are then defined based on the original query combined with the cluster names. New expanded queries are used to search for new results. Each independent expanded query is then treated as a semantically well-defined cluster of images. A user case study has been realized with 24 volunteers and the results show a preference over two different clustered versions i.e. image retrieval and theme retrieval. Unfortunately, as they use the ephemeral clustering algorithm of [Zeng et al., 2004], they show similar drawbacks as in [Ding et al., 2008].

Even with these efforts, the most common problems of text cluster retrieval are also present in image cluster retrieval. Recently, [Dias et al., 2011] described a way to obtain hierarchical clusters over Web snippets. The results show that this technique is a recent state-of-the-art technique in ephemeral clustering, however, the results are not completely compared to others ephemeral clustering algorithms. In Section 2.2.2, two text-based approaches are explored avoiding visual information. In particular, we propose the use of two different approaches: (1) a similar framework as proposed in [Wang et al., 2007] but using the algorithm presented in [Dias et al., 2011] and (2) a query log-based query expansion technique. These two approaches are particularly relevant to mobile Web image retrieval as only a few clusters of semantically related images can be retrieved and as a consequence provide an effortless interface for mobile devices as browsing will be intrinsically limited.

2.2.2 Text-based vs Query Log Clustering by Expansion

Our exploration consists in two different methodologies to address the organization of Web image search results. First, an ephemeral clustering approach is presented and a second one is based on query logs. Independently of the approach, the proposed algorithms must achieve the following goal: building a Web image search results taxonomy

for mobile devices in response to a text query in order to facilitate user interaction and exploration. Moreover, the methodologies will have to be based on commercial search engines services to retrieve Web images, Web snippets and query suggestions, so that we provide an image meta search engine for mobile devices capable to deal with real-world queries.

As previous works in text document retrieval [Zeng et al., 2004] [Carpineto and Romano, 2009], we propose to use a recent search results clustering algorithm which allows to fit the image search results on a small screen of a mobile device⁷. In [Dias et al., 2011], the authors address the additional challenges of (1) finding the suitable label for each cluster, which can be different from words contained in a cluster and (2) proposing a method that finds only a few number of clusters. In contradistinction with classical search results clustering, our idea is to group a collection of images that comes from the Web in response to a text query. For that purpose, we propose to expand any text query following two different approaches. With the ephemeral clustering approach, the query is expanded with each cluster name to form a cluster of related images. With the query log based approach, we use the user query to suggest new query expansion to form different groups of similar pictures.

Mobile Interface

An interface for mobile devices has been developed using the Android platform for touch screen devices⁸. An initial start page is displayed on which the user can enter the text query and the Web image search results are displayed in groups of images as shown in Figure 2.2 for the query “jaguar”. The left side corresponds to the ephemeral clustering based strategy and the right side to the query log approach. This is a typical query used to check the ability of the system to find semantically separated results. In particular, both approaches retrieve semantically separated results. The ephemeral clustering based approach finds the cluster names “sells services”, “cars new”, “land rover” and “onca america” (the animal) for the first four image groups and the query log based approach suggests “jaguar”, “jaguar usa”, “jaguar car” and “jaguar animal”. To facilitate the exploration phase, each group of image results can be explored with left and right movements as a typical gallery exploration. Moreover, the different groups can be explored using up and down movements allowing a quick exploration of different meanings involved in the original query. Finally, when the user clicks an image, he gets the basic information of the image and with a longer click, the user is redirected to the image website.

⁷Note that any search results clustering algorithm can be used in this phase. However, we chose the one proposed by [Dias et al., 2011] for facility reasons.

⁸In this part, the interface is quickly presented. Further details about the user interaction capabilities are included in Section 5.2.1

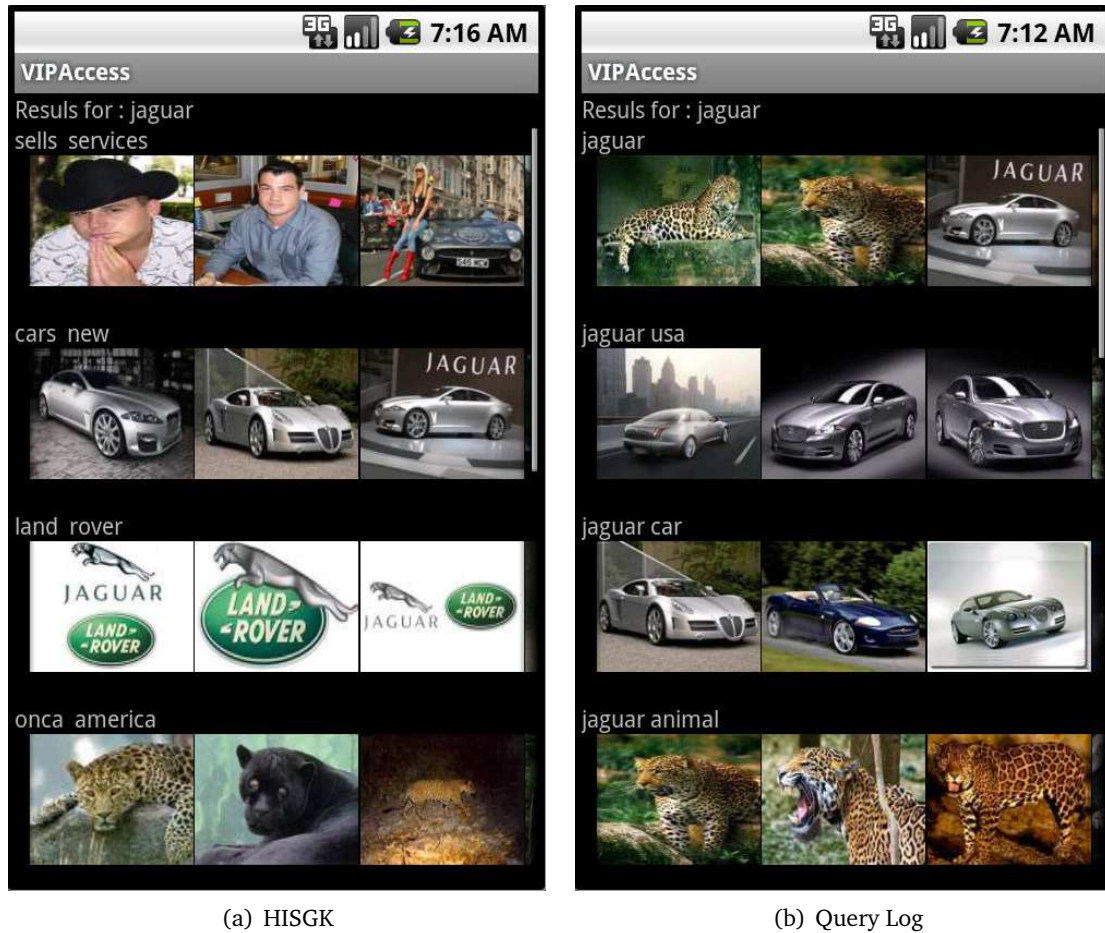


Figure 2.2: Results for the query "jaguar" using Ephemeral Clustering (HISGK-means [Dias et al., 2011]) and Query Logs.

Ephemeral Clustering Based Approach

The ephemeral clustering algorithm proposed by [Dias et al., 2011] has been used for this clustering approach. This algorithm uses Web snippet results to generate an automatic hierarchical representation. The clusters are groups of Web page results and are labeled during the clustering process. The procedure used to get Web image search clusters by expansion is defined in Algorithm 2.1.

The results are displayed on the mobile device as sorted groups, as shown in Figure 2.2. Each group is the returned set of images obtained by the Image Search API from Google when the string *ExpandQuery* from step 5 (Algorithm 2.1) is used as a query, following the cluster by expansion technique.

Algorithm 2.1 Web image clustering by expansion using search results clustering algorithms.

Input: *TextQuery* **Output:** *ImagesCluster*

1. *RankedList* = *getWebResults(TextQuery)*
2. *ClusterSet* = *SearchResultsClustering(RankedList)*
3. For each element *cluster_i* in *ClusterSet*
4. *ClusterName* = *getClusterName(cluster_i)*
5. *ExpandQuery* = *concat(TextQuery, ClusterName)*
6. *ImagesCluster_i* = *getImageResults(ExpandQuery)*
7. *ImageClusterName_i* = *ClusterName*
8. return *ImagesCluster*

Query Log Based Approach

This approach is based on query logs obtained by Google Suggestions API [Google, 2011a]. As far as we know, there are no similar previous studies to organize image Web search topics, even though previous studies have been developed for organizing Web search results [Wang and Zhai, 2007]. The Google Suggestions API service allows to obtain frequently-used queries related to any given query. Word completion is a particular procedure, which allows the user to be aware of queries constructed by users using the original query as a prefix. For example, when the query “app” is used in the Google Suggestions API, the top suggestions are “applebees”, “apple store”, “apple trailers” and so on. These queries do not in fact accurately represent the intentions of the original query. However, if we add a blank space, we get suggestions such as “app store”, “app world”, “app planet” which are more related to the original query. In this previous approach, a Web image search is done with the query expansion and the results are displayed in the same way as in the first approach. Algorithm 2.2 describes the query log based approach.

Algorithm 2.2 Image clusters with Query Logs.

Input: *TextQuery* **Output:** *ImagesCluster*

1. *TextQuery* = *concat(TextQuery, blankSpace)*
2. *QueriesSet* = *getQueryLogSuggestions(TextQuery)*
3. For each element *QueriesSet_i* in *QueriesSet*
4. *ExpandQuery* = *concat(TextQuery, QueriesSet_i)*
5. *ImagesCluster_i* = *getImageResults(ExpandQuery)*
6. *ImageClusterName_i* = *ExpandQuery*
7. return *ImagesCluster*

2.2.3 Evaluating Clustering by Expansion

In this preliminary experiment, evaluation is the challenging part. Many metrics have been proposed for clustering evaluation, however, our technique of query by expansion

does not follow classical clustering ideas and as a consequence, the use of classical evaluation metrics is not feasible. For this purpose, to verify the performance of both approaches i.e. ephemeral clustering and query logs, a user evaluation is proposed. In this initial evaluation, the entire Web image content is used as a data set by Google Search API. Moreover, to avoid the subjectivity of biased user evaluations, we propose to evaluate the results using the Amazon crowd-sourcing platform called Mechanical Turk⁹ (AMTurk). With respect to the query set used to evaluate the methodologies, we chose the top fastest rising and top falling queries from 2010 provided by Google Zeitgeist 2010¹⁰. Some of these queries were removed as they contained straight characters¹¹. Finally, 97 queries were used to retrieve images and evaluate the approaches. In the 11 categories in Table 2.1, the number of queries from each category and their respective queries are shown. In the last column, the original query (in bold font) and examples of query expansions found by both approaches are shown. These queries were used in both experiments, by the users and automatic evaluations.

Users Evaluation

Using the selected queries and described approaches, we retrieved the top four relevant Web images of the three most relevant clusters. Cluster relevance order was defined by using the estimation of the number of results for each query returned by the image search API. Both results (i.e. ephemeral clustering and query log) are presented to three different users using AMTurk, who have to choose between the following options: (1) the left side is better, (2) both are bad, (3) both are good or (4) the right side is better. To avoid any evaluation bias due to Web image positioning, the images were presented randomly on the right or left side, independently of the used methodology. Three different situations have to be considered by the evaluators to assess the results: (1) taking into just the text label account, (2) taking just the Web image results into account and finally, (3) taking both text label and Web image results into account. An agreement phase is then done to avoid unusual judgments in the post-processing task of the collected data. Each query was evaluated by three judges and when at least two out of three judges agreed, the answer was taken into account, otherwise the judgment was ignored. The user agreement obtained was 78%, i.e., at least two users assigned the same answer to 75 out of the 97 original queries. The overall results and corresponding discussion are presented in Section 2.2.4.

⁹<https://www.mturk.com/mturk/welcome> [Lask access: 20/06/2014].

¹⁰<http://www.google.com/intl/en/press/zeitgeist2010/> [Lask access: 20/06/2014].

¹¹Of course the cluster by expansion technique could deal with these cases, but to avoid not English character, query drop was necessary.

Category	#	Queries	Expanded Queries Examples
Fastest Rising	10	chatroulette, ipad, justin bieber, nicki minaj, friv, myxer, katy perry, twitter, gamezer, facebook	chatroulette : 2010 new, alternative, alternative for adults, ban, chat video, chat world, clones, code script, gifs, ip blocked, like sites, not working, random website, screenshots, service know, service new, sites, source code
Fastest Falling	10	swine flu, wamu, new moon, mininova, susan boyle, slumdog millionaire, circuit city, myspace layouts, michael jackson, national city bank	
Entertainment	8	shakira, eminem, netflix, youtube videos, lady gaga, kesha, groove-shark, transformers 3	shakira : 1977 born, albums biography, biography, downloads songs, facebook official, loca, loca lyrics, lyrics, news lyrics, pictures gallery, rabiosa, rabiosa lyrics, waka waka
Sports	10	mundial 2010, olympics, espn3, fifa 11, randy moss, miami heat, mourinho, wayne rooney, cricket live score, david villa	
Consumer Electronics	8	iphone 4, nokia 5530, htc evo 4g, nokia n900, blackberry apps, duracell mygrid, otterbox, pdanet	iphone 4 : 2011 phone, 3gs apple, accessories, apple 2010, apps, cases, covers, jailbreak, recording calling, reviews, specs, unlock, verizon
Food & Drink	8	masterchef, cupcakes, jimmy johns, dominos pizza menu, tudo gostoso receitas, guacamole recipe, applebees menu, rachel ray	
Maps Searches	10	anhembi parque, wm gucken, world cup, bundeskanzleramt, rio branco, mt everest, kew gardens, tour eiffel, oxford street, nürburgring	anhembi parque : america latin, brazil pavilion, complex shows, inn holiday, paulo brazil, paulo turismo, sao paulo, travel youtube
People	6	selena gomez, kim kardashian, miley cyrus, taylor lautner, megan fox, robert pattinson	
News	7	haiti, besiktas, chile, earthquake, jörg kachelmann, mobile technology, oil spill	haiti : earthquake, earthquake facts, earthquake struck, economy finance, election, flag, history, libre, map, news, news information, president, weather, western country, world development
Health Queries	10	hcg diet, dr oz, aspergers, mcdonalds nutrition, vitamin d deficiency, appendicitis symptoms, cholera, nfp, vacina h1n1, whooping cough	
Humanitarian Aid	10	donate to haiti, donate to pakistan, text to donate, doctors without borders, download to donate, red cross canada, blood donation restrictions, donate blood australia, donate now button, csl plasma	donate to haiti : 2010 relief, donations earthquake, earthquake, help earthquake, hope now, red cross, red cross, relief, text

Table 2.1: Categorization, number and queries used for the evaluations.

Automatic Evaluation

To automatically evaluate the coverage of the expanded results, we propose and analyze two different factors. First, we propose a new evaluation measure called the Average Accumulative Estimated Results (AccER). The AccER is calculated using the estimated size of retrieved results for each query, eventually expanded. The objective of the AccER measure is to define how many images are retrieved accumulatively with r query expansions. The AccER is defined in Equation 2.1 where $S_{1j}, S_{2j}, \dots, S_{rj}, \forall j = 1, \dots, n$ are the estimated sizes of image results for the different expanded queries 1, 2, ..., r of the original query j , knowing that n original queries exist.

$$AccER_r = \sum_{j=1}^n S_{rj} + \sum_{k=1}^{r-1} AccER_k \quad (2.1)$$

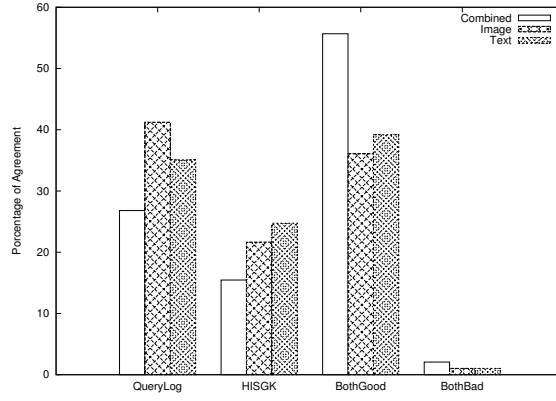


Figure 2.3: Percentage of users agreement for Text, Images and Combined questions from 45 AMTurk workers.

Second, to evaluate if the expanded queries adequately overlap, a percentage of this overlapping is calculated from the original query. The overlapping percentage is defined in Equation 2.2 where R_{q_1} is the result set of retrieved images for the query q_1 and $R_{q_i} \forall i = 1, \dots, r$ are the result sets for each expansion of the query q_1 . In particular, r is the number of expansions obtained and m is the maximum size of the real retrieved images (in our experiments $m = 64$).

$$\%Overlapping = \sum_{j=1}^n \frac{\sum_{i=1}^r size(R_{q_j^i} \cap R_{q_j})}{r * m} * 100 \quad (2.2)$$

Higher values of this metric indicate a better response of overlapping between the original query and the expanded queries and can be interpreted as the average percentage of images in one expanded query, which were retrieved by the original query. Results and discussion are presented in the next Section 2.2.4.

2.2.4 Results and Discussion

Users Evaluation

A total of 45 workers were involved in the users evaluation phase through AMTurk. They solved 291 tasks including three times each one of the 97 queries in Table 2.1. The results about the preference of the users are shown in Figure 2.3. Each percentage corresponds to one of the situations presented to the user and their respective clustering by expansion technique. These percentages correspond to the results obtained under the user agreement rule as described above.

Figure 2.3 shows the distribution results for Image, Text and Combined (Image and Text) judgments. The results show a higher percentage of agreement value for the answer

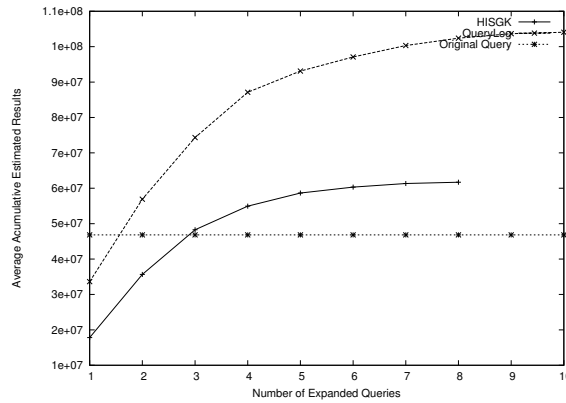


Figure 2.4: Estimated Average Accumulative results over the 97 queries for the first 10 image groups.

“both are good”. This result encourages the continuation of our research in this kind of visualization. Comparatively, results for the query log based expansion obtain higher acceptance rates when compared to the ephemeral clustering counterpart. As a result, users tend to understand the results shown better with the query log approach rather than the ephemeral clustering strategy. Nevertheless, it is important to note that the queries used were issued in 2010, which can explain the good performance obtained by the query log approach. However, the ephemeral clustering approach receives a good rating from users as well since both methodologies are accepted as good by the users. Moreover, it tends to show cluster names better than the query log methodology comparing to the accuracy of image clusters. These results show a promising field to explore when both approaches can be combined in a mixed strategy that may explore the best of each one. On the one hand, the ephemeral clustering approach is a better option when the query is rarely used and it is not possible to find good query logs. On the other hand, the query log approach is a good option when the queries are frequent ones.

Automatic Evaluation

The results of automatic evaluation are shown in Figure 2.4. This allows us to check the average accumulative estimated results (AccER) for the original query and the expansion obtained with both strategies.

As the curve of the ephemeral clustering approach has a better approximation than the original query curve, this can be interpreted as a better generation of new query names without overlapping, while query log based estimated results show that this approach includes a high number of overlapped results¹². This factor is important in the moment of determining the extra redundant results included in each cluster. This issue is particularly

¹²Note that in this case the interpretation of overlapping corresponds to the overlapped Web results of the query expansions and not to the classical interpretation of overlapping as the possibility of one document to belong to many topics.

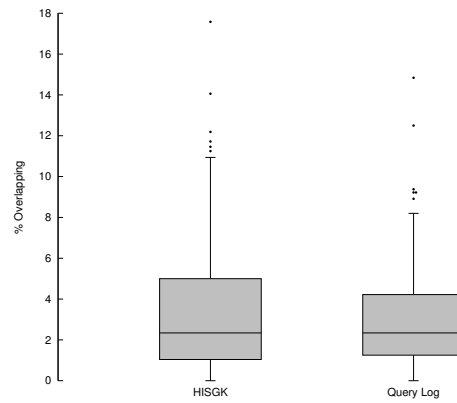


Figure 2.5: Boxplot for percentage of overlapping between the first 64 results of the original query and the first 64 results of the expanded queries.

important for image exploration, as more overlapping images immediately decreases the user interface success.

Additionally, in order to check the overlapping between the original query and the new expanded queries, the percentage of overlapping was calculated and the results for the 97 queries are shown in Figure 2.5. The results show similar behaviors in both approaches and the paired t-student test indicates that there is not a significant difference between the two approaches.

With these evaluations, user-based and automatically-based, the obtained results confirm in a consistent way that both the proposed approaches perform well in the organization of Web image results. Furthermore, the two seem to be complementary.

Although these short experiments allow us to draw out the possibilities of expansion by clustering techniques for Web image results exploration in mobile devices, the automatic evaluation is not a classical way to evaluate ephemeral clustering algorithms. For this reason, a more clustering oriented evaluation is presented in the remainder of this chapter. Indeed, more classical search results clustering techniques are studied to identify the capabilities of well-known clustering techniques in this specific context.

2.3 Ephemeral Clustering and their baselines

As afore mentioned, visualization of Web search results is a challenging problem in Information Retrieval (IR). Previous works have been addressing this problem in many interesting ways. For example, in order to deal with ambiguous or multifaceted queries, Web results are grouped in correlated content instead of long flat lists of relevant documents. Among existing techniques, ephemeral clustering¹³ is a commonly studied area, which consists in providing clusters of Web re-

¹³Also called Web Search Results Clustering (SRC).

sults generated “on-the-fly” and based on their content, typically short texts or Web snippets. Many works have been presented including task adapted clustering [Moreno et al., 2013], meta clustering [Carpineto and Romano, 2010] and knowledge-based clustering [Scaiella et al., 2012][Moreno et al., 2014]¹⁴.

However, the evaluation of clustered data is a hot topic for related areas of Natural Language Processing (NLP) and IR. For example, in IR, a recent study received much attention by comparing well-known clustering evaluation metrics [Amigó et al., 2009]. In their conclusion, the authors show the importance of choosing adequate metrics and recommend the use of the F_{b3} -measure. Within the specific case of SRC, different metrics have been used such as F_1 -measure (F_1), $kSSL$ ¹⁵ and F_{b3} -measure (F_{b3}). And even worse, possible evaluations between clustered results and references could return different values of F -measure. Consequently, no standard evaluation framework has been adopted by the SRC community. The generation of SRC datasets is a hard task as well. In fact, many works evaluate their results over different standard datasets: ODP-239 [Carpineto and Romano, 2010] and Moresque [Navigli and Crisafulli, 2010]. Unfortunately, comparative results are usually biased as baseline algorithms are run over default parameter configurations whereas proposed methodologies are usually tuned to increase their performance over the studied metrics and/or datasets.

In this section, we focus on deep understanding of the evaluation task within the context of SRC. First, we provide the results of baseline algorithms with their best parameter settings through a full exploration using exhaustive combination of datasets and metrics. Second, we show that a simple cascade strategy of baseline algorithms can lead to a scalable and real-world solution, which evidences comparative results to recent text-based algorithms. Finally, we draw some conclusions about evaluation metrics and an interesting particularity in SRC, their bias to the number of output clusters.

2.3.1 Search Results Clustering and Related Work

Search results clustering has been an active area of research during the last two decades. The exponential growth of Internet and new available resources as well as clustering techniques have motivated many works in this area. Figure 2.6 presents the main SRC techniques developed over time. In fact, two main streams can clearly be recognized: text-based strategies such as [Cutting et al., 1992, Hearst and Pedersen, 1996, Zamir and Etzioni, 1998, Zeng et al., 2004, Osinski et al., 2004, Carpineto and Romano, 2010, Carpineto et al., 2011, Moreno et al., 2013] and knowledge-based ones [Ferragina and Gulli, 2008, Scaiella et al., 2012, Di Marco and Navigli, 2013, Moreno et al., 2014]. The first group

¹⁴Note that some of these works are presented in this thesis. So, the main discussion will be presented in further Chapters.

¹⁵This subjective metric is used for labeling evaluation. This topic will be addressed in Section 4.5.

concentrates on proposing news algorithms based on string matching, frequency properties or clusters agreement. The latter use external resources to improve one or more of the typical clustering stages. For example, word occurrences in alternative collections, such as Wikipedia [Scaiella et al., 2012] or Google N-grams [Di Marco and Navigli, 2013], can be used to improve similarities between them. However, additional computation is necessary in these cases.

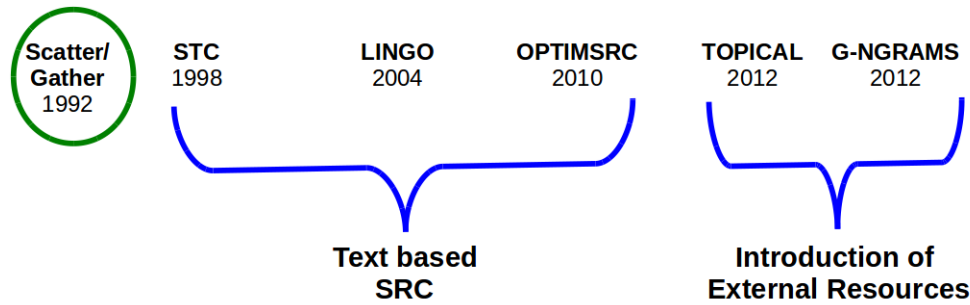


Figure 2.6: Representative works of SRC over time: STC [Zamir and Etzioni, 1998], LINGO [Osinski and Weiss, 2005], OPTIMSRC [Carpineto and Romano, 2010], TOPICAL [Scaiella et al., 2012] and G-NGRAMS [Di Marco and Navigli, 2013].

Often when SRC comparison is necessary, Suffix Tree Clustering (STC) [Zamir and Etzioni, 1998] and LINGO [Osinski et al., 2004] are used. These two SRC algorithms provide publicly available libraries and as a consequence, are often used as state-of-the-art baselines. Indeed, many recent studies mainly compare with LINGO and STC due their well-studied capabilities. However, they are frequently underestimated because of their parameters being not adequately defined or explored. Main SRC techniques are presented below including text-based and knowledge-based ones.

STC: [Zamir and Etzioni, 1998] defined the Suffix Tree Clustering algorithm which is still a difficult standard to beat in this field. The main peculiarity of this algorithm is that it was defined to achieve simplicity and efficacy. In particular, they propose a monothetic clustering technique which merges documents based on string overlap. Indeed, instead of using the classical Vector Space Model (VSM) representation, they propose representing Web snippets using compact tries and clusters which are defined by the documents with shared roots. Note that the more overlapping is present in the string characters, the more related the documents are and the more probability there is for them to be assigned to the same cluster.

LINGO: [Osinski and Weiss, 2005] proposed a polythetic solution called LINGO, which takes the string representation proposed by [Zamir and Etzioni, 1998] into account. First, common phrases are extracted by the use of suffix-arrays. The term-document matrix is then built using word frequencies and weights are calculated using the standard “term

frequency-inverse document frequency” (tfidf) technique. Then, to discover documents with latent structures, the Single Value Decomposition (SVD), similar to latent semantic indexing [Deerwester et al., 1990], is applied with the matrix. Finally, latent dimensions are considered as extracted latent topics and the relevant documents are assigned to them. LINGO tends to assign documents to the “Other” cluster when the belongs threshold value is not achieved.

BiKm: [Steinbach et al., 2000] presented bisecting K-means (BiKm) as a simple algorithm in which the set of documents is iteratively partitioned in two parts until achieve the desire number of clusters. Finally, clusters are labeled only with individual words found in the Web snippets of each cluster, but may not be present in all the documents in the cluster.

OPTIMSRC: [Carpineto and Romano, 2010] showed that the complementary characteristics of the SRC algorithm’s outputs, as LINGO and STC, suggest the adoption of the meta clustering approach. For this purpose, they introduce a novel criterion to measure the concordance of two partitions of Web snippets into different clusters based on the information content associated with the series of decisions made by the partitions on the single pairs of Web snippets. A meta clustering phase is then casted to an optimization problem of the concordance between the clustering combination and the given set of clusterings. The results of their OPTIMSRC system demonstrate that meta clustering is superior over individual clustering techniques. Additionally, they propose a dataset called ODP-239, which is widely used in the community.

TOPICAL: [Scaiella et al., 2012] proposed a top performing SRC system over the ODP-239 dataset. They propose to move away from the bag of words representation towards a graph of topics paradigm derived from TAGME, a wikification algorithm [Vitale et al., 2012]. Each Web snippet is annotated with a set of topics, which are represented by Wikipedia articles. A bipartite-like graph structure is built where nodes are either Web snippets or topics and edges are either topic-to-topic or topic-to-snippet. Then, a spectral-like clustering algorithm is run over the graph to discover relevant clusters and meaningful labels. TOPICAL is an interesting approach as clustering is driven by the presence of Wikipedia titles in Web snippets and somehow assures the quality of the labeling.

G-NGRAM: [Di Marco and Navigli, 2013] recently proposed a SRC algorithm which relies on Web n-grams. The first step- in order to capture the similarity better between Web snippets- consists in building a co-occurrence graph based on Dice coefficients calculated over the Google Web1T corpus [Brants and F., 2006] from which meanings are discovered by word meaning induction algorithms. Each Web snippet is represented as a bag of words

(polythetic approach) but their similarity is computed by discovered word senses. Their experiments show that enhanced diversification and clustering performance results can be obtained based on the adjusted RandIndex [Vinh et al., 2009] for a specific dataset built for ambiguous queries (Moresque). Recently, researchers from the same team proposed a new dataset within the context of the SEMEVAL task 11 [Navigli and Vannella, 2013], whose goal is to provide an evaluation framework for the objective comparison of word sense disambiguation and induction algorithms in SRC for ambiguous queries.

2.3.2 Evaluation Metrics and Datasets

Evaluation Metrics

Different metrics have been proposed to evaluate text clustering. Within this study, we present the most relevant metrics used in dealing with the SRC. The first complete study in terms of evaluation has certainly been proposed by [Carpineto and Romano, 2010]. The authors propose the use of the F_1^C metric¹⁶, which is a specific implementation of the more general F_β measure. Other metrics have also been proposed in alternative studies. For example, the F_{b3} measure [Amigó et al., 2009] addresses many important problems in clustering such as cluster homogeneity, completeness, rag-bag and size-vs-quantity constraints, and has shown interesting properties for the SRC task as formulated in [Moreno et al., 2013]. Two other important metrics have been studied in [Di Marco and Navigli, 2013]: F_1^N and the Adjusted RandIndex (ARI) [Vinh et al., 2009]. F_1^N can be seen as a complementary metric of F_1^C as it is also based on the classical F_β measure but is computed in a different manner¹⁷, whereas ARI evidences a good property for the SRC. While it measures clustering accuracy, it also takes into account the fact that a given partition shows a similar partitioning shape compared to the reference gold standard. The underlying idea is that the number of clusters and the average number of Web snippets in each cluster approximate the reference clustering as close as possible. An illustration of this situation can be seen in [Moreno and Dias, 2013b]¹⁸. In Table 2.2, we've defined all the metrics used in our experiments. Note that not all the metrics are used in this Chapter. However, in order to have a unique notation we define the metrics used in SRC evaluation here¹⁹.

¹⁶In order to identify each evaluation metric, we use to different notations for F_1 the F_1^C and F_1^N .

¹⁷Let us notice that these are two F_1 measures whose computation is defined differently in [Carpineto and Romano, 2010] and [Di Marco and Navigli, 2013].

¹⁸For more details refer to Chapter 5.

¹⁹Most of the chapters contain at least one of the mentioned evaluation metrics.

Evaluation Metric	where
$F_{\beta}^C = \frac{(1+\beta^2)*P*R}{\beta P+R}$	$P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN}, TP = \sum_{i=1}^k \sum_{x_j \in \pi_i^*} \sum_{x_l \in \pi_i^*, l \neq j} g_0(x_j, x_l), FP = \sum_{i=1}^k \sum_{x_j \in \pi_i} \sum_{x_l \in \pi_i, l \neq j} (1 - g_0^*(x_j, x_l)),$ $FN = \sum_{i=1}^k \sum_{x_j \in \pi_i^*} \sum_{x_l \in \pi_i^*, l \neq j} (1 - g_0(x_j, x_l))$
$F_{b^3} = \frac{2*P_{b^3}*R_{b^3}}{P_{b^3}+R_{b^3}}$	$P_{b^3} = \frac{1}{N} \sum_{i=1}^k \sum_{x_j \in \pi_i} \frac{1}{ \pi_i } \sum_{x_l \in \pi_i} g_0^*(x_j, x_l), R_{b^3} = \frac{1}{N} \sum_{i=1}^k \sum_{x_j \in \pi_i^*} \frac{1}{ \pi_i^* } \sum_{x_l \in \pi_i^*} g_0(x_j, x_l)$
$F_1^N = \frac{2*P*R}{P+R}$	$P = \frac{1}{\sum_{i=1}^k \pi_i } \sum_{i=1}^k \arg \max_{\pi_z^*} \left(\sum_{x_j \in \pi_z^*} \sum_{x_l \in \pi_i} g_1(x_j, x_l) \right), R = \frac{1}{\sum_{i=1}^k \pi_i^* } \sum_{i=1}^k \sum_{x_j \in \pi_i^*} \sum_{p=1}^k \sum_{x_l \in \pi_p} g_1(x_j, x_l)$
$ARI(\Pi, \Pi^*)$	$ARI(\Pi, \Pi^*) = \frac{RI(\Pi, \Pi^*) - E(RI(\Pi, \Pi^*))}{\max RI(\Pi, \Pi^*) - E(RI(\Pi, \Pi^*))} \text{ where}$ $RI(\Pi, \Pi^*) = \frac{TP+TN}{TP+FP+FN+TN}, TN = N - TP - FP - FN$
and	$g_0(x_i, x_j) = \begin{cases} 1 & \iff \exists l : x_i \in \pi_l \wedge x_j \in \pi_l \\ 0, & otherwise \end{cases} \quad g_0^*(x_i, x_j) = \begin{cases} 1 & \iff \exists l : x_i \in \pi_l^* \wedge x_j \in \pi_l^* \\ 0, & otherwise \end{cases}$ <p>where π_i is the cluster solution i ($\Pi = \cup \pi_i$) and π_i^* is the gold standard of the category i ($\Pi^* = \cup \pi_i^*$).</p> $g_1(x_i, x_j) = \begin{cases} 1 & \iff x_i = x_j \\ 0, & otherwise \end{cases}$

Table 2.2: Clustering Evaluation Metrics.

SRC Datasets

Different gold standards have been used for the evaluation of SRC algorithms, among which the most cited are²⁰: ODP-239²¹ [Carpineto and Romano, 2010], Moresque²² [Navigli and Crisafulli, 2010] and SEMEVAL²³ [Navigli and Vannella, 2013]. Following is described the process used to built each of these datasets.

ODP-239 In this dataset, each document is represented by a title and a Web snippet and the subtopics are chosen from the top levels of DMOZ²⁴. In this dataset, the number of possible subtopics is always equal to 10. Each query, over 239 of them, corresponds to one category present in DMOZ and as Web results were used the Web sites associated to the respective category. For each Web result, the Web site description from DMOZ is used as Web snippet because they share similar size. Usually the Web snippet description is limited to 25-30 words.

Moresque This dataset was built using 114 queries that correspond to disambiguation pages of Wikipedia. As such, these subtopics are likely to cover most of the senses present in the Web results. For each query, the top 100 Web results returned by the Yahoo! SE were tagged with the most appropriate query senses found in the disambiguation pages of Wikipedia.

SEMEVAL This dataset was similarly built as Moresque, e.g. the subtopics are defined based on the disambiguation pages of Wikipedia. However, Web results are obtained using another commercial SE, the Google SE. The rest of the annotation process in the dataset was performed following the procedure defined for Moresque.

A quick summary of both datasets is presented in Table 2.3. Note that in average, Moresque includes less subtopics than SEMEVAL. There are four queries²⁵ for which the number of subtopics is equal to 1. Indeed, the SEMEVAL dataset could be seen as a cleaner version of Moresque. For this reason, in the remainder of this chapter we use the Moresque dataset, but in the following chapters we present our results using ODP-239 and SEMEVAL.

²⁰Other SRC dataset is AMBIENT [Bernardini et al., 2009], but it is a less complete version of Moresque.

²¹<http://credo.fub.it/odp239/> [Last acc.: Jan., 2014]

²²<http://lcl.uniroma1.it/moresque/> [Last acc.: Jan., 2014]

²³<http://www.cs.york.ac.uk/semeval-2013/task11/> [Last acc.: Jan., 2014]

²⁴<http://www.dmoz.org> [Last acc.: Jan., 2014]. DMOZ is also known as the Open Directory Project (ODP).

²⁵Queries id number: 46, 68, 86 and 111.

Dataset	Number of queries	Number of Subtopics by query			Number of Web snippets
		Average	Minimum	Maximum	
ODP-239	239	10	10	10	25580
Moresque	114	3.7	1	11	11350
SEMEVAL	100	7.7	2	19	6400

Table 2.3: Description of the SRC gold standard datasets.

2.3.3 Current Results

As afore mentioned, a good number of studies have addressed the SRC problem. However, comparing all of them is a tedious task due to the fact that no standard evaluation has been adopted. As a consequence, all works have used different datasets or evaluation measures. Particularly, the well-known F_1 has been used more frequently as an evaluation metric. More recently, [Carpineto and Romano, 2010] evidenced more complete results of the general definition of the F_β -measure for $\beta = \{1, 2, 5\}$, [Navigli and Crisafulli, 2010] introduced the Rand Index metric and [Moreno et al., 2013] used F_{b3} (introduced by [Amigó et al., 2009]) for a more adequate clustering metric.

In Table 2.4²⁶, we report the results obtained so far in the literature by text-based and knowledge-based strategies for the F_1^C metric with ODP-239 and Moresque datasets. Results show that knowledge-based strategies performs better than text-based. However, in some cases these differences are not only because the knowledge-based algorithms performs better, but also because text-based algorithms are not adequately configured ²⁷.

		F_1^C	
		ODP-239	Moresque
Text	STC	0.324	0.455
	LINGO	0.273	0.326
	OPTIMSRC [Carpineto and Romano, 2010]	0.313	-
	[Moreno et al., 2013]	0.390	0.665
Know.	TOPICAL [Scaiella et al., 2012]	0.413	-
	G-NGRAM [Di Marco and Navigli, 2013]	-	0.7204*

Table 2.4: State-of-the-art results using F_1^C for several SRC algorithms. (*) See table footnote.

Note that, evaluation results are only available for the ODP-239 dataset using the F_1^C metric, but not for the other metrics. For example, the results for TOPICAL [Scaiella et al., 2012] not include the values obtained when other β values different to 1 are used, i.e. $\beta = 2$ and $\beta = 5$. And similarly, G-NGRAM [Di Marco and Navigli, 2013] does not provided results for ODP-239 neither for other existing dataset, but only for a

²⁶The result of [Di Marco and Navigli, 2013] is based on a reduced version of AMBIENT + Moresque as authors do not include results for individual datasets and they cannot be derived.

²⁷This will be discussed in Section 2.3.4.

new combination of two of them. We experiment in deep with existing implementations of commonly used SRC baselines in the next section. Our aim is to use a deal of recent clustering metrics with most common datasets to achieve a better understanding of the text-based baselines algorithms.

2.3.4 Understanding Baseline Algorithms and Metrics

When a newly proposed algorithm is presented, it is usually tuned to its maximal performance. However, the results of baseline algorithms are usually run with their default parameters based on available implementations. Therefore, no conclusive remarks can be drawn knowing that tuned versions might provide improved results.

Publicly available implementations²⁸ of STC and LINGO include a fixed stopping criterion. However, it is well-known that tuning the number of output clusters may greatly impact the clustering performance in almost any clustering task. Indeed, discovering the appropriate number of clusters is considered a challenging task. However, in order to provide fairer results for frequently used baseline algorithms, we evaluated a k -dependent²⁹ version for the implemented baselines. For that reason, we ran both algorithms for $k = 2..20$ and chose the best result as the “optimal” performance in an ideal case where the algorithm knows the correct number of desired clusters. Table 2.6 sums up the results for all the baselines in their different configurations and shows that tuned versions outperform standard (available) ones— both for F_1^C and F_{b^3} over ODP-239 and Moresque.

Moresque								
Algo.	F_1^C				F_{b^3}			
	Stand.	k	Tuned	k	Stand.	k	Tuned	k
STC	0.4550	12.7	0.6000	2.9	0.4602	12.7	0.4987	2.9
LINGO	0.3258	26.7	0.6034	3.0	0.3989	26.7	0.5004	5.8

Table 2.5: Standard, Tuned and Random Results for Moresque dataset.

OPD-239								
Algo.	F_1^C				F_{b^3}			
	Stand.	k	Tuned	k	Stand.	k	Tuned	k
STC	0.3238	12.4	0.3350	3.0	0.4027	12.4	0.4046	14.5
LINGO	0.2029	27.7	0.3320	3.0	0.3461	27.7	0.4459	8.7

Table 2.6: Standard, Tuned and Random Results for ODP-239 dataset.

²⁸<http://carrot2.org> [Last access: 19/06/2014]

²⁹Carrot2 parameters *maxClusters* and *desiredClusterCountBase* are used to set k value for STC and LINGO, respectively.

		Moresque					
		F_1^C			F_{b3}		
Level 1	Level 2	Stand.	Equiv.	k	Stand.	Equiv.	k
STC	STC	0.6145	0.5594	3.1	0.4550	0.4913	3.1
	LINGO	0.5611	0.4932	7.3	0.4980	0.4716	7.3
LINGO	STC	0.5696	0.5176	6.7	0.4602	0.4854	6.7
	LINGO	0.4629	0.4371	13.7	0.4447	0.4566	13.7

Table 2.7: Cascade Results for Moresque datasets.

Fair Configuration of Baseline Algorithms

In the previous section, our aim was to claim that tunable versions of existing baseline algorithms might evidence improved results when faced with the ones reported in the literature. These values should be taken as the “real” baseline results in controllable environments. However, exploring all of the parameter space is not an applicable solution in a real-world situation where the reference is unknown. Therefore, a stopping criterion must be defined to adapt to any dataset distribution. This is the particular case in the standard implementations of STC and LINGO.

Previous results [Carpineto and Romano, 2010] showed that different SRC algorithms provide different results and hopefully complementary ones. For instance, STC demonstrates high recall and low precision, while LINGO inversely evidences high precision for low recall. Iteratively applying baseline SRC algorithms may thus lead to improved results by exploiting each algorithm’s strengths.

In a cascade strategy, we first cluster the initial set of Web page snippets with any SRC algorithm. Then, the input of the second SRC algorithm is the set of meta-documents built from the documents belonging to the same cluster³⁰. Finally, each clustered meta-document is mapped to the original documents generating the final clusters. This process can be iteratively applied, although we only take two-level cascade strategies into account in this thesis³¹.

This strategy could be viewed as an easy, reproducible and parameter free baseline SRC implementation that should be compared to existing state-of-the-art algorithms. Tables 2.7 and 2.8 show the results obtained with different combinations of SRC baseline algorithms for the cascade strategy of both F_1^C and F_{b3} over ODP-239 and Moresque. The “Stand.” column corresponds to the performance of the cascade strategy and k to the automatically obtained number of clusters. Results show that the combination STC-STC achieves the best overall performance of F_1^C , and STC-LINGO is the best combination of F_{b3} in both datasets.

³⁰Fused using concatenation of strings.

³¹More than two-level were evaluated but, because the results were not improved, the values are not shown.

		ODP-239					
		F_1^C			F_{b3}		
Level 1	Level 2	Stand.	Equiv.	k	Stand.	Equiv.	k
STC	STC	0.3629	0.3304	3.2	0.3982	0.4023	3.2
	LINGO	0.3624	0.3258	6.9	0.4249	0.4010	6.9
LINGO	STC	0.3457	0.3029	7.2	0.4229	0.4429	7.2
	LINGO	0.2789	0.2690	13.6	0.3931	0.4237	13.6

Table 2.8: Cascade Results for ODP-239 datasets.

In order to provide a more complete evaluation, in column “Equiv.” we included the performance that could be obtained by the tunable version of each single baseline algorithm based on the same k . Interestingly, the cascade strategy outperforms the tunable version of any k for F_1^C but fails (not by far) to compete with F_{b3} . This issue is discussed in the next section.

Metric Bias

In Table 2.6 , one can see that when using the tuned version and evaluating with F_1^C , the best performance for each baseline algorithm is obtained for the same number of output clusters as independently of the dataset (i.e. around 3 for STC and LINGO). Therefore, the quick conclusion would be that the tuned versions of STC and LINGO are strong baselines as they show similar behaviour over datasets. In a realistic situation, k might be directly tuned to this value.

However, when comparing the output number of clusters based on the best F_1^C value to the reference number of clusters, a huge difference is evidenced. Indeed, in Moresque, the ground-truth average number of clusters is 6.6 and exactly 10 in ODP-239. Interestingly, F_{b3} shows more accurate values for the number of output clusters in the best tuned baseline performances. In particular, the best F_{b3} results are obtained by LINGO with 5.8 clusters for Moresque and 8.7 clusters for ODP-239 which approximate the ground-truths the most.

In order to better understand the behaviour of each evaluation metric (i.e. F_β^C and F_{b3}) over different k values, we experienced a uniform random clustering over Moresque and ODP-239. A uniform random clustering algorithm randomly assigns documents to each cluster in a way that, in the end, each partition includes equally sized clusters. In Figure 2.7, we illustrate these results. The important issue is that F_β^C is more sensitive to the number of output clusters than F_{b3} . On the one hand, all F_β^C measures provide the best results for $k = 2$. Furthermore, a random algorithm could reach $F_1^C=0.5043$ for Moresque and $F_1^C=0.2980$ for ODP-239, thus outperforming almost all standard implementations of STC and LINGO for both datasets. On the other hand, F_{b3} shows that most standard baseline implementations outperform the random algorithm. Henceforth, our

analysis will consider more relevant the clustering evaluation results obtained with F_{b3} than with other metrics.

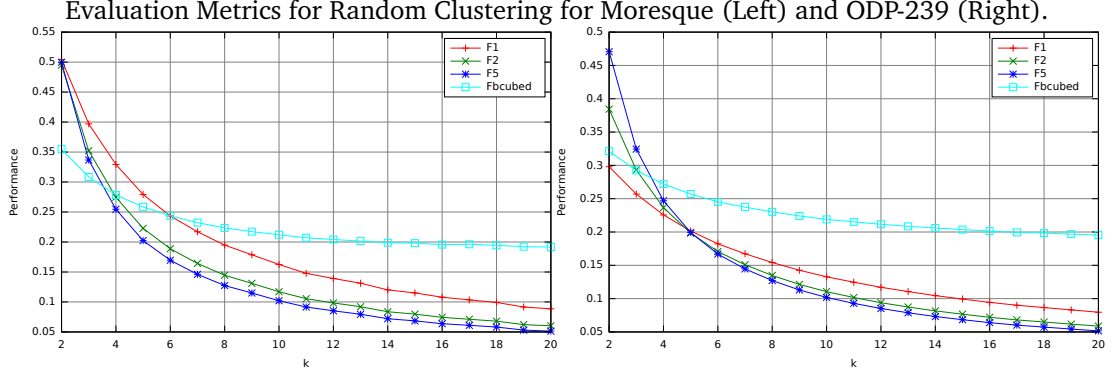


Figure 2.7: F_{b3}^C and F_{b3} for Moresque and ODP-239 for Random Clustering.

In Figures 2.8(a) and 2.8(b), we illustrate the different behaviours of F_1^C and F_{b3} for $k = 2..20$ for both standard and tuned versions of STC, LINGO and BiKm. One may clearly see that F_{b3} is capable of discarding the algorithm (BiKm) that performs worst in the standard version, whereas this is not the case for F_1^C . For LINGO, the optimal performances over Moresque and ODP-239 are near the ground-truth number of clusters while this is not the case for F_1^C which evidences a decreasing tendency when k increases.

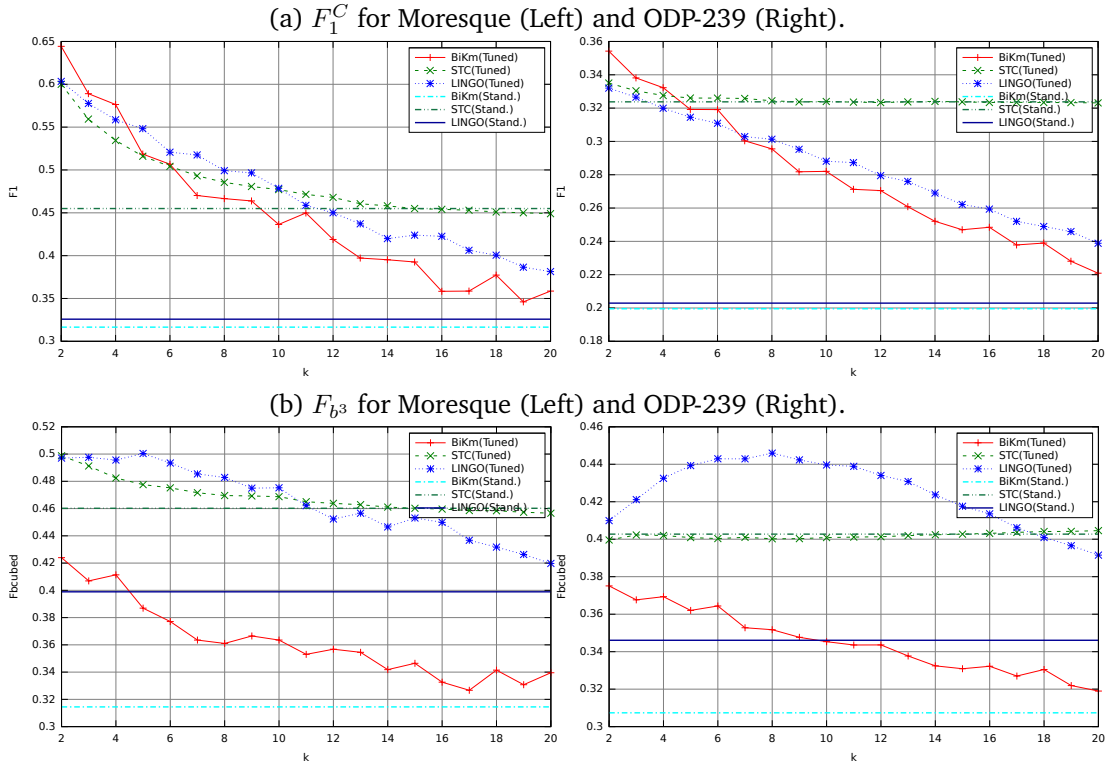


Figure 2.8: F_1^C and F_{b3} for Moresque and ODP-239 for Standard and Tuned Clustering.

In section 2.3.4, we showed that competitive results could be achieved with a cascade strategy based on baseline algorithms. Although results outperform standard and tun-

able baseline implementations for F_1 , it is wise to use F_{b3} to better evaluate the SRC task, according to our previous discussion. In this case, the best values are obtained by STC-LINGO with $F_{b3}=0.4980$ for Moresque and $F_{b3}=0.4249$ for ODP-239, which highly approximate the values reported in [Moreno et al., 2013]: $F_{b3}=0.490$ (Moresque) and $F_{b3}=0.452$ (ODP-239). Additionally, when STC is performed first and LINGO after, the cascade algorithm scales better. This can be explained because LINGO does not scale as well as it does STC³². Indeed, as LINGO is based on a matrix decomposition operation the time consumption grows exponentially with the number of snippets to cluster.

2.4 Conclusions

In this Chapter, we present our knowledge about the state-of-the-art techniques. A complete analysis is presented about two main topics: mobile image Web retrieval and text search results clustering. First, a user study is performed so as to understand their impressions in terms of Web image results visualization and label quality for a given set of results. We perform a two fold evaluation, user-based and automatic evaluations have shown that the proposed visual paradigm performs well in a real-world environment when popular queries are used. Indeed, the ephemeral clustering technique shows more compact results when compared with the query log-based technique. Nevertheless, the query log-based technique evidences better user acceptance rates in terms of labeling quality. Both solutions have already been implemented into a new combination of common gallery interfaces that facilitates the results exploration task of organized image Web results by semantic groups.

Next, an in-depth study about the cluster quality is performed to achieve the state-of-the-art capacity in its domain. This part presents a discussion about the use of baseline algorithms in SRC and evaluation metrics. Our experiments show that F_{b3} seems more adapted to evaluate SRC systems than the commonly used F_1^C with the standard datasets available so far. For this reason, in the next chapters our conclusions will be more influenced by the results obtained with F_{b3} than results of F_1^C . However, when clustering quality is evaluated, other metrics (F_β^C , F_1^N and ARI) will be included for comparison reasons and to identify other behaviors.

Additionally, we presented a new baseline strategy which approximate state-of-the-art algorithms in terms of clustering performance and that can be obtained by an easy, reproducible and parameter free implementation. It will be considered as the hard baseline result for the rest of this thesis.

In the following of this thesis, our efforts will be directed towards the developing of new SRC algorithms capable of integrating a well-performing algorithm in terms of cluster

³²<http://carrotsearch.com/lingo3g-comparison> [Last access: 19/06/2014]

quality and labeling quality. Our first developed algorithm will be presented in the next chapter, the *AGK*-means algorithm, which integrates in a unique phase the clustering and labeling, two commonly separated steps.

Publications

This chapter has been validated by the following publications [Moreno and Dias, 2014a, Moreno and Dias, 2011]:

Moreno, J. G. and Dias, G. (2011). Using ephemeral clustering and query logs to organize web image search results on mobile devices. In Proceedings of International ACM workshop on Interactive Multimedia on Mobile and Portable Devices (IMMPD), pages 33–38.

Moreno, J. G. and Dias, G. (2014). Easy web search results clustering: When baselines can reach state-of-the-art algorithms. In Proceedings of 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pages 1–5.

Chapter 3

Improving Text-based Ephemeral Clustering

Contents

2.1 Introduction	17
2.2 Web Image Retrieval on Mobile Devices	18
2.2.1 Text-based Web Image Clustered Visualization	21
2.2.2 Text-based vs Query Log Clustering by Expansion	22
2.2.3 Evaluating Clustering by Expansion	25
2.2.4 Results and Discussion	28
2.3 Ephemeral Clustering and their baselines	30
2.3.1 Search Results Clustering and Related Work	31
2.3.2 Evaluation Metrics and Datasets	34
2.3.3 Current Results	37
2.3.4 Understanding Baseline Algorithms and Metrics	38
2.4 Conclusions	42

3.1 Introduction

In SRC, for any given query, the retrieved Web snippets are automatically clustered and presented to the user with meaningful labels in order to minimize the information searching process. This technique can be particularly useful for polysemous queries but it is hard to implement efficiently and effectively [Carpineto et al., 2009]. Indeed, as opposed to classical text clustering, SRC must deal with small collections of short text fragments (Web snippets) and process them in run-time.

As mentioned in Section 2.3.1, most of the successful methodologies follow a monothetic approach [Zamir and Etzioni, 1998, Ferragina and Gulli, 2008, Carpineto and Romano, 2010, Navigli and Crisafulli, 2010, Scaiella et al., 2012]. The underlying idea is to discover the discriminant topical words in the collection and group together Web snippets containing these relevant terms. On the other hand, the polythetic approach, whose main idea is to represent Web snippets as word feature vectors has received less attention, the only relevant study being [Osinski and Weiss, 2005].

The main reasons for this situation are:

- (i) word feature vectors are hard to define in small collections of short text fragments [Timonen, 2013],
- (ii) existing similarity measures, such as the cosine, are unadapted to capture the semantic similarity between small texts,
- (iii) corpus-based techniques, such as latent semantic analysis, have evidenced inconclusive results [Osinski and Weiss, 2005] or are impractical for SRC and
- (iv) the labeling process is a surprisingly hard extra task [Carpineto et al., 2009] when classical text clustering techniques are applied.

In this Chapter, we explore the introduction of a polythetic approach to improved existing techniques in SRC when it is correctly applied to small collections of short text fragments or Web snippets. For this purpose, we propose a new methodology that adapts a classical algorithm, particularly the K -means algorithm, in order to access a third-order similarity measure based on well-known collocation measures and initially developed out of the SRC community [Dias et al., 2007]. Moreover, our adapted version of the K -means algorithm allows the labeling of each cluster directly within the clustering process - thus avoiding the above - mentioned extra task. In the same fashion, collocation measures are introduced into the labeling task as a measurement of similarity between labels and Web snippets. Finally, the evolution of the objective function of the adapted K -means version is modeled to automatically define the number of clusters.

As proof of this concept, we propose different experiments with the ODP-239 [Carpineto and Romano, 2010] and Moresque [Navigli and Crisafulli, 2010] datasets against the most competitive text-based SRC algorithms studied in Section 2.3: STC [Zamir and Etzioni, 1998], LINGO [Osinski and Weiss, 2005], OPTIMSRC [Carpineto and Romano, 2010] and the classical bisecting incremental K -means (which may be seen as a baseline for the polythetic paradigm)¹. Moreover, we use the evaluation measure F_{b3} [Amigó et al., 2009] to evaluate both cluster homogeneity and completeness. Results evidence that our proposal outperforms strong text-based state-of-the-art

¹In this Chapter, we purposely omit TOPICAL [Scaiella et al., 2012] and G-NGRAM [Di Marco and Navigli, 2013] algorithms for comparison since both of them are knowledge-driven methodologies and they will be studied in Chapter 4.

approaches with a maximum $F_{b3} = 0.452$ for ODP-239 and $F_{b3} = 0.490$ for Moresque.

3.2 Polythetic Search Results Clustering

3.2.1 Intuitive Idea

Before formally introducing our method, its underlying idea will be briefly described. First, it is important to remark that performing labeling and clustering in an unified algorithm is a challenging task. Consider for example the classical K -means algorithm. It is a well-known strategy for clustering. However, in its classical implementation, labels are not provided. We could propose a naïve solution that provides labels using the centroid representations after the clustering process. However, there are two main drawbacks: first, similarity between short texts using classical vector representations tends to be zero so that all documents are mutually close; and second, labels could easily overlap generating clusters with identical labels. To overcome these problems, we propose to represent documents and centroids using context vectors. A context vector is not a vector of values but rather a vector of words. The use of context vectors facilitates the labeling process due to the fact that final centroids are used as labels. However, to produce comprehensible labels, they must remain short and informative. To do this, a small number p is used to indicate the maximum number of words used in the context vector. Then, each document must be represented with an under sampled of p words² since expected labels are normally shorter than documents and large values of p increase the label complexity. For the final step, a context vector clustering algorithm must be applied. Consequently, new issues must be addressed. In the next Section, we present our clustering algorithm based on the classical K -means algorithm and third-order similarity metrics.

3.2.2 Polythetic Clustering

The classical version of K -means is a geometric clustering algorithm [Lloyd, 1982]. Given a set of n data points, the algorithm uses a local search approach to find a local optimal partition of the points in a given number of K clusters. A set of initial K cluster centers are then chosen to start the process. Next, each point is assigned to the center closest to it. Finally, the centers are recomputed as centers of mass of their assigned points. This process is repeated until a desired grade of convergence is achieved. To assure convergence³, an objective function J_K is defined which is minimized (or maximized) at each processing step. In the classical version of K -means, this objective function is defined as in Equation 3.1 where π_k is the cluster k ($\Pi = \{\pi_1, \dots, \pi_k\}$), $s_i \in \pi_k$ is a data

²The specific method to select the p words is presented below.

³In the local search.

point in the cluster π_k , m_{π_k} is the centroid of the cluster π_k ($M = \{m_1, \dots, m_k\}$) and $E(.,.)$ is the euclidean distance.

$$J_K(\Pi, M) = \sum_{k=1}^K \sum_{s_i \in \pi_k} E(s_i, m_{\pi_k})^2. \quad (3.1)$$

In the case of SRC, the original version of the K -means algorithm is not suitable⁴. The biggest drawback is related to sparseness in short texts. Web snippets need to be highly informative in a constrained space. For this purpose, only a few sentences or words are displayed in which the most relevant information is concentrated. Of course, related documents could use alternative words to refer to the same subject. In this case, the euclidean distance or cosine similarity could fail due the missing overlap in the vector dimensions.

Therefore, classical K -means could be adapted to integrate more suitable similarity measures [Mihalcea et al., 2006]. Particularly, we propose the use of third-order similarity metrics that exploit capabilities of collocation measures. Third-order similarity measures (also called weighted second-order similarity measures) do not rely on exact matches of word features as classical measures/distances (e.g. cosine similarity or euclidean distance), but rather evaluate similarity based on related matches using, e.g., collocation measures. We propose to use a simplified version of a third-order similarity measure introduced in [Dias et al., 2007] and defined in Equation 3.2.

$$S_{3rd}(s_i, s_j) = \frac{1}{p^2} \sum_{k=1}^p \sum_{l=1}^p s^{w_{ik}} * s^{w_{jl}} * C(w_{ik}, w_{jl}). \quad (3.2)$$

Given two Web snippets $s_i = \{w_{i1}, \dots, w_{ip}\}$ and $s_j = \{w_{j1}, \dots, w_{jp}\}$, their similarity is evaluated by the similarity of its constituents based on a collocation metric $C(.,.)$ ⁵ where w_{ik} (resp. w_{jl}) corresponds to the word at the k^{th} (resp. l^{th}) position in the vector s_i (resp. s_j) and $s^{w_{ik}}$ (resp. $s^{w_{jl}}$) is the weight of word w_{ik} (resp. w_{jl}) in the set of retrieved Web snippets. Accordingly, the J_K objective function must be adapted to the S_{3rd} similarity measure. A direct consequence is the definition of a new objective function $J_{S_{3rd}}$ that must be maximized⁶. This function is defined in Equation 3.3.

$$J_{S_{3rd}} = \sum_{k=1}^K \sum_{s_i \in \pi_k} S_{3rd}(s_i, m_{\pi_k}). \quad (3.3)$$

⁴Of course, it is still possible to apply the K -means algorithm, but it usually underperforms when compared to other algorithms.

⁵It is important to realize that $C(.,.)$ is a value based on word distribution over each Web snippet retrieved for a given query. These values are used during the entire clustering process and recalculated when a new Web snippet set is provided.

⁶A maximization process can easily be transformed into a minimization one.

Note that, given the restriction of the number of constituents p for each element compared by S_{3rd} a new procedure to define s_i and m_{π_k} is needed. Indeed, if p is inferior to the number of different words in the Web snippet only the most relevant p words are used to represent it. For each Web snippet, an interestingness value is calculated for each word belonging to it and the p words with higher values are selected. Interestingness values for Web snippets are calculated using the $\lambda^s(w)$ function defined in Equation 3.4.

$$\lambda^s(w) = \sum_{w_q^i \in s_i} C(w_q^i, w), \forall w \in s_i. \quad (3.4)$$

Correspondingly, a cluster centroid m_{π_k} is similarly defined as a Web snippet, i.e., by a context vector of p words $(w_1^{\pi_k}, \dots, w_p^{\pi_k})$. As a consequence, each cluster centroid must be instantiated in such a way that $J_{S_{3rd}}$ increases at each step of the clustering process. The choice of the best p words representing each cluster is our implicit way of labeling each cluster. For this purpose, we define a label composition procedure which consists in selecting the best p words from the global vocabulary V in such a way that $J_{S_{3rd}}$ increases for each step. The global vocabulary V is the set of all the words which appear in the initial set of retrieved Web snippets.

So, for each word $w \in V$ and a given collocation metric $C(., .)$, its interestingness $\lambda^k(w)$ is computed in regards to cluster π_k . With this function, the interestingness of each word is evaluated. This operation is defined in Equation 3.5 where $s_i \in \pi_k$ is a Web snippet from cluster π_k . Finally, the p words with higher $\lambda^k(w)$ are selected to construct the cluster centroid and used as cluster labels in such a way that, $J_{S_{3rd}}$ is locally maximized for each step, similarly to how it is performed in the classical K -means. Note that a word may not be part of the representation used for its respective Web snippet but, under our strategy, even when a word is not part of the cluster π_k it may be part of the centroid m_{π_k} ⁷.

$$\lambda^k(w) = \frac{1}{p} \sum_{s_i \in \pi_k} \sum_{w_q^i \in s_i} S(w_q^i, w). \quad (3.5)$$

In order to provide an automatic definition of the number of k clusters, we propose to rely on a modified version of the K -means algorithm called Adapted Global K -means (AGK-means) [Likasa et al., 2003], which has proved to lead to improved results. To solve a clustering problem with M clusters, all intermediate problems with $1, 2, \dots, M - 1$ clusters must be sequentially solved. In this strategy, the underlying idea is that an optimal solution for a clustering problem with M clusters can be obtained using a series of local searches using the K -means algorithm. In each local search, the $M - 1$ cluster centers are initially always placed at their optimal positions corresponding to the clustering problem with $M - 1$ clusters. The remaining M^{th} cluster center is initially placed at several posi-

⁷This interesting effect is given by the full search over the global vocabulary V

tions within the data space. In addition to effectiveness, the method is deterministic and does not depend on any initial conditions or empirically adjustable parameters. Moreover, its adaptation to the SRC is straightforward. In Section 3.3, the complete consideration for the automatic selection of k is presented.

3.2.3 Collocation Measures

Collocation measures, also called association measures, evaluate the degree of association between candidates of collocation [Pecina, 2005]. In recent years, collocation measures have become widely used in many applications such as machine translation, word sense disambiguation, language generation, and information retrieval [Downey et al., 2007]. A collocation function usually receives a pair of words as input and returns a number indicating if the pair of words occurs together more often than by chance. However, these functions have a preprocessing step in which distribution of frequencies are analyzed over a corpus. In the SRC, this corpus corresponds to the set of Web snippets retrieved for one query.

Note that Equations 3.2, 3.3, 3.4 and 3.5 are defined in terms of collocation measures. In the actual experiments, two association measures which are known to have different behaviors [Pecina and Schlesinger, 2006] are studied. We implement the Symmetric Conditional Probability [Silva et al., 1999] in Equation 3.6 which tends to give more credit to frequent associations and the Point wise Mutual Information [Church and Hanks, 1990] in Equation 3.7, which over-estimates infrequent associations. Note that w_{ik} and w_{jl} are two words contained in the context vectors.

$$SCP(w_{ik}, w_{jl}) = \frac{P(w_{ik}, w_{jl})^2}{P(w_{ik}) \times P(w_{jl})}. \quad (3.6)$$

$$PMI(w_{ik}, w_{jl}) = \log_2 \frac{P(w_{ik}, w_{jl})}{P(w_{ik}) \times P(w_{jl})}. \quad (3.7)$$

3.3 Stopping Criterion

Once the clustering has been processed, selecting the best number of clusters still remains to be done. For this, numerous procedures have been proposed [Milligan and Cooper, 1985]. However, given that the proposed algorithm does not perform as the usual K -means, none of the listed methods is directly adaptable to our specific solution. To overcome this situation, we propose a specific procedure based on the definition of a rational function which adapts to our objective function $J_{S_{3rd}}$ and models the quality of each partition obtained with the previously described Adapted Global K -means

strategy. Certainly, an ideal candidate number of clusters can be found if each intermediate partition of M clusters is evaluated under the proposed quality criterion. Therefore, each time that an intermediate partition is obtained with the Adapted Global K -means strategy, the $J_{S_{3rd}}$ value is calculated and stored. Then, consecutive values are analyzed to select the ideal candidate. To better understand the behavior of $J_{S_{3rd}}$, at each step of the AGK -means algorithm we present its values for a given example when $k = 2..10$ in Figure 3.1.

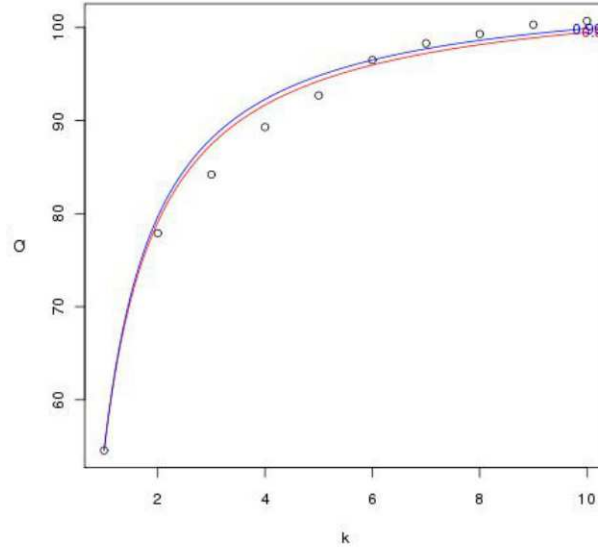


Figure 3.1: $J_{S_{3rd}}$ and its modelisation.

$J_{S_{3rd}}$ can be modeled as in Equation 3.8 which converges to a limit α when K increases and starts from $J_{S_{3rd}}^1$ (i.e. $J_{S_{3rd}}$ at $K = 1$). This situation is embodied by the blue line in Figure 3.1, i.e. the line which intersect the dot at intersection $K = 6$. The idea to discover the optimal K , is to chose the β value which maximizes the difference from the average β^{mean} . So, α , β and γ need to be expressed independently of unknown variables.

$$\forall K, f(K) = \alpha - \frac{\gamma}{K^\beta}. \quad (3.8)$$

Using Equation 3.8, we can find an alternative definition for γ by only replacing the $K = 1$ and $f(1) = J_{S_{3rd}}^1$.

$$J_{S_{3rd}}^1 = \alpha - \frac{\gamma}{1^\beta}. \quad (3.9)$$

Since α can operationally be defined and $\gamma = \alpha - J_{S_{3rd}}^1$, then β can be defined in terms of only α . By replacing the γ factor from Equation 3.9 in Equation 3.8 and by definition

$f(K) = J_{S_{rd}}^K$, we obtain:

$$J_{S_{rd}}^K = \alpha - \frac{\alpha - J_{S_{rd}}^1}{K^\beta}, \quad (3.10)$$

where $J_{S_{rd}}^K$ is the value of the Equation 3.3 when evaluated for K . Finally, after some basic operations, we obtain β defined by the Equation 3.11.

$$\beta = \frac{\log(\alpha - J_{S_{rd}}^1) - \log(\alpha - J_{S_{rd}}^K)}{\log(K)}. \quad (3.11)$$

Now, the value of α which best approximates the limit of the rational function must be defined. For that purpose, we computed its maximum theoretical and experimental values as well as its approximated maximum experimental value based on the δ^2 -Aitken [Aitken, 1926] procedure to accelerate convergence as explained in [Kuroda et al., 2008]. The best results were obtained with the maximum experimental value which is defined by building the cluster centroid m_{π_k} for each Web snippet individually. Finally, the best number of clusters is calculated as shown in Algorithm 3.1 and each cluster is given its label based on the p words with greater interestingness of its centroid m_{π_k} .

Algorithm 3.1 The best K selection procedure.

1. Calculate β^K for each K
 2. Evaluate the mean of all β^K i.e. β^{mean}
 3. Select β^K which maximizes $\beta^K - \beta^{mean}$
 4. Return K as the best number of partitions
-

This situation is illustrated in Figure 3.1 where the red line corresponds to the rational functional for β^{mean} and the blue line models the best β value (i.e. the one which maximizes the difference from β^{mean}). In this case, the best number would correspond to β^6 and as a consequence, the best number of clusters is 6. In order to illustrate the soundness of the procedure, we present the different values of β at each K iteration and the differences between consecutive values of β at each iteration in Figure 3.2. We clearly see that the highest inclination in the curve is between cluster 5 and 6 which also corresponds to the highest difference between two consecutive values of β .

3.4 Text-based Evaluation

Evaluating SRC systems is a difficult task as stated in [Carpineto et al., 2009]. Indeed, an evaluation setup for SRC systems must consider mainly two aspects: label quality and cluster quality. In this Chapter, we will concentrate our efforts on the latter. Ideally, each manually defined query subtopic should be matched with one and only one

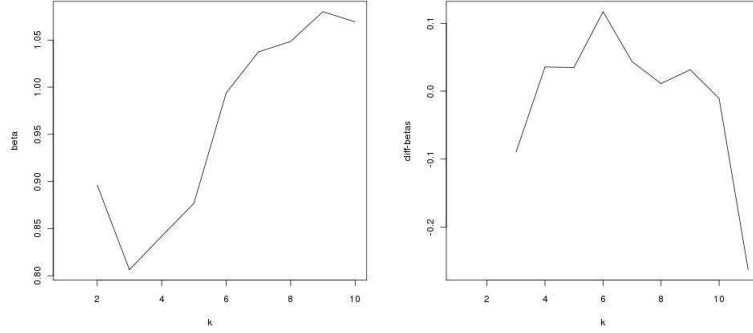


Figure 3.2: Values of β (on the left) and differences between consecutive values of β (on the right).

cluster obtained by the algorithm to be evaluated. However, so far this task is far from being achievable. Indeed, actual evaluation metrics consider different ways to compare the cluster results with the reference, thus metrics based on individual, pair or cluster matching are used.

Certainly, measuring small differences in terms of the cluster quality of two similar but different partitions is a challenging task [Amigó et al., 2009]. As such, this constraint is reformulated as follows: the task of SRC systems is to provide complete topical cluster coverage of given queries, while avoiding excessive redundancy of the subtopics in the result list of clusters. So, in order to evaluate our clustering algorithm, as well as the cluster size determination, we propose two different evaluations setups. First, we want to evidence the quality of the stopping criterion when compared to an exhaustive exploration of all possible K values in a given interval. Secondly, we propose a comparative evaluation of our algorithm with existing text-based state-of-the-art algorithms with gold standard datasets and recent clustering evaluation metrics⁸.

3.4.1 Text Processing

Before the clustering process takes place, Web snippets are represented as word feature vectors and collocation measure values are calculated for each individual query. The first step consists in determining the words probabilities used in Equations 3.6 and 3.7. Then, these values are used to provide the collocation function response when requested. This preprocessing step is performed using the Web service proposed in [Dias et al., 2011]⁹.

Next, a relevance score is assigned to any word present in the set of retrieved Web snippets based on the analysis of the context words through the use of the interestingness $\lambda^s(w)$

⁸Specifically, F_β^C and $F_{b,3}$ are analyzed.

⁹Access to this Web service is available upon request to the authors.

function defined in Equation 3.4. Then, each Web snippet is represented by the set of its p most relevant in the sense of the $\lambda^s(w)$ function.

Finally, once the preprocessing task is finished, the clustering task is performed as mentioned in previous Sections.

3.4.2 Intrinsic Evaluation

The first set of experiments focuses on understanding the behavior of our methodology within a greedy search strategy for different tunable parameters defined as a tuple $\langle p, K, C(w_{ik}, w_{jl}) \rangle$. In particular, p is the size of the word feature vectors representing both Web snippets and centroids (varying from 2 to 5), K is the number of clusters to be found (varying from 2 to 10) and $C(w_{ik}, w_{jl})$ is the collocation measure integrated in the third order similarity measure. In these experiments, two collocation measures were selected *SCP* [Silva et al., 1999] (Equation 3.6) and *PMI* [Church and Hanks, 1990] (Equation 3.7)¹⁰. Then, the best obtained $\langle p, K, C(w_{ik}, w_{jl}) \rangle$ configurations are compared to the proposed stopping criterion.

In order to perform this task, we evaluate performance based on the F_{b3} measure introduced in [Amigó et al., 2009] over the ODP-239 gold standard dataset proposed in [Carpineto and Romano, 2010]. In particular, [Amigó et al., 2009] indicate that common metrics such as the F_β -measures are good for assigning higher scores to clusters with high homogeneity, but fail to evaluate cluster completeness. The first results are provided in Tables 3.1 and 3.2. The results evidence that the best configurations for different $\langle p, K, C(w_{ik}, w_{jl}) \rangle$ tuples are obtained for high values of p , K ranging from 4 to 6 clusters and *PMI*, in Table 3.2, steadily improving over *SCP*, in Table 3.1. However, additional experiments must be performed to analyze the stopping criterion. As such, our proposed stopping strategy evidences coherent results because

- (i) it does not depend on the used association measure since the results obtained by *SCP* and *PMI* are similar, $F_{b3}^{SCP} = 0.452$ and $F_{b3}^{PMI} = 0.450$,
- (ii) it automatically discovers similar numbers of clusters independently of the size of the p -context vector used to represent the Web snippets or clusters, and
- (iii) performance is increased when higher values of p are used.

3.4.3 Comparative Evaluation

The second evaluation aims to compare our methodology to the current state-of-the-art text-based SRC algorithms. Similar to in Section 2.3, we propose comparative

¹⁰Note that other collocation measures can be used. However, the study of other measures is performed in the next Chapter.

		p			
		2	3	4	5
K	2	0.387	0.400	0.405	0.408
	3	0.396	0.411	0.416	0.422
	4	0.398	0.412	0.423	0.431
	5	0.396	0.409	0.425	0.431
	6	0.391	0.406	0.423	0.429
	7	0.386	0.400	0.420	0.429
	8	0.382	0.397	0.416	0.423
	9	0.378	0.391	0.414	0.422
	10	0.374	0.388	0.411	0.421
	Stop	F_{b^3}	0.395	0.411	0.452
Criterion	Avg. K	4.799	4.690	4.766	4.778

Table 3.1: F_{b^3} for SCP for the global search and the stopping criterion for the ODP-239 dataset.

		p			
		2	3	4	5
K	2	0.391	0.408	0.420	0.423
	3	0.399	0.418	0.434	0.444
	4	0.397	0.422	0.439	0.451
	5	0.393	0.418	0.439	0.451
	6	0.388	0.414	0.435	0.451
	7	0.383	0.410	0.430	0.445
	8	0.377	0.405	0.425	0.441
	9	0.373	0.398	0.420	0.434
	10	0.366	0.392	0.412	0.429
	Stop	F_{b^3}	0.393	0.416	0.450
Criterion	Avg. K	4.778	4.879	4.874	4.778

Table 3.2: F_{b^3} for PMI for the global search and the stopping criterion for the ODP-239 dataset.

experiments for two gold standard datasets (ODP-239 [Carpineto and Romano, 2010] and Moresque [Di Marco and Navigli, 2013]) for STC [Zamir and Etzioni, 1998], LINGO [Osinski and Weiss, 2005], OPTIMSRC [Carpineto and Romano, 2010] and the strong baseline strategy STC-LINGO presented in Section 2.3.4¹¹. For OPTIMSRC, we reproduced the results presented in the paper of [Carpineto and Romano, 2010] as no implementation is freely available.

Results for the Moresque dataset are presented in Table 3.4¹². Our strategy outperforms all the baseline algorithms in terms of F_{β}^C . Regarding the F_{b^3} , our strategy is close to outperforms almost all them, except for STC-LINGO configuration of the cascade algorithm which slightly beats our strategy. However, for all configurations of p and $C(., .)$, our al-

¹¹A complete description of the baselines is given in Section 2.3.1.

¹²Note that results for the OPTIMSRC are not available for the Moresque dataset.

		F_1^C	F_2^C	F_5^C	F_{b3}
SCP	2	0.312	0.363	0.411	0.395
	3	0.341	0.393	0.441	0.411
	4	0.352	0.404	0.453	0.441
	5	0.366	0.416	0.462	0.452
PMI	2	0.332	0.363	0.390	0.393
	3	0.358	0.395	0.430	0.416
	4	0.378	0.421	0.459	0.436
	5	0.390	0.435	0.476	0.450
STC [Zamir and Etzioni, 1998]		0.324	0.319	0.322	0.403
LINGO [Osinski and Weiss, 2005]		0.273	0.167	0.153	0.346
STC-LINGO [Moreno and Dias, 2014a]		0.362	N/A	N/A	0.425
OPTIMSRC [Carpineto and Romano, 2010]		0.313	0.341	0.380	N/A

Table 3.3: SRC comparative results for F_β^C and F_{b3} over the ODP-239 dataset.

gorithm achieves state-of-the-art results. Indeed, our worst result is $F_{b3} = 0.462$ and it is not far from our best: $F_{b3} = 0.490$, which indicates the robustness of the algorithm. On the other hand, it is clear that our strategy outperforms baselines for the ODP-239 dataset presented in Table 3.3. Indeed, similar properties of robustness can be observed in the results also.

Another interesting issue is related to the p value. As in intrinsic evaluation, the higher the p value, the better the performance. This phenomena is supported by the F_β^C results where the situation is similar. For design reasons, higher p values increase the number of words used to describe the cluster, and as a consequence, increase the difficulty of label understanding. Note that, labeling evaluation is out of the scope of this Chapter. Also clustering evaluation results indicate that higher p values clearly increase clustering performance. In the following chapter, we explore the possibility of integrating this into a more general algorithm.

Finally, contrary to the results obtained with ODP-239 when evaluated with F_{b3} , Moresque is sensitive to the selection of collocation measures. This situation can be explained by the characteristics of Moresque. This dataset was built to evaluate SRC systems and it is one of the first attempts to use Web results obtained from a commercial search engine. However, annotation of the obtained results is a difficult task. In the available dataset, their authors did not include labels when disagreement was presented by the annotators since many documents are unannotated. This situation clearly affects the results when F_{b3} is used. We strongly believe that these missing annotations make the drawing of correct conclusions from this dataset difficult and recommend the use of a new version introduced by the same authors in [Navigli and Vannella, 2013]¹³.

¹³Our following experiments use this new version.

		F_1^C	F_2^C	F_5^C	F_{b3}
SCP	2	0.627	0.685	0.747	0.482
	3	0.649	0.733	0.817	0.482
	4	0.665	0.767	0.865	0.473
	5	0.664	0.770	0.872	0.464
PMI	2	0.615	0.644	0.679	0.490
	3	0.551	0.548	0.563	0.465
	4	0.543	0.521	0.519	0.462
	5	0.571	0.551	0.553	0.485
STC [Zamir and Etzioni, 1998]		0.455	0.392	0.370	0.460
LINGO [Osinski and Weiss, 2005]		0.326	0.260	0.237	0.399
STC-LINGO [Moreno and Dias, 2014a]		0.5611	N/A	N/A	0.498

Table 3.4: SRC comparative results for F_β^C and F_{b3} over the Moresque dataset.

3.5 Conclusions

In this Chapter, we proposed a new text-based SRC approach which (1) is based on the adaptation of the K -means algorithm to third-order similarity measures and (2) proposes a coherent stopping criterion. Results evidenced clear improvements over the evaluated state-of-the-art text-based approaches as well as over the strong baselines¹⁴ for two gold standard datasets. Moreover, our best F_1^C -measure over ODP-239 (0.390) approximates the highest ever-reached F_1^C -measure (0.413) by the TOPICAL knowledge-driven algorithm proposed in [Scaiella et al., 2012]¹⁵.

Furthermore, the analysis performed over p value indicates that higher values achieve better performance, although it could mean additional efforts in label understanding when a user explores the final clustered results¹⁶. To remedy this situation, an adequate strategy must deal with higher p values while avoiding extra efforts in label understanding. Indeed, the results are promising and in the next chapter of this theses, we are going to define a new knowledge-based third-order similarity measure based on external resources, which would propose the beginning of an answer to this problem.

Publications

This chapter has been validated by the following publication [Moreno et al., 2013]:

Moreno, J. G., Dias, G., and Cleuziou, G. (2013). Post-retrieval clustering using third-order similarity measures. In Proceedings of 51st Annual Meeting of the Association for Computational Linguistics (ACL), pages 153–158.

¹⁴Like those presented in previous Chapter.

¹⁵Notice that the authors only propose the F_1^C -measure although different results can be obtained for different F_β^C -measures and F_{b3} as evidenced in Tables 3.3 and 3.4.

¹⁶Note that evaluating this situation is a big challenge.

Chapter 4

Including External Information in Ephemeral Clustering

Contents

3.1	Introduction	45
3.2	Polythetic Search Results Clustering	47
3.2.1	Intuitive Idea	47
3.2.2	Polythetic Clustering	47
3.2.3	Collocation Measures	50
3.3	Stopping Criterion	50
3.4	Text-based Evaluation	52
3.4.1	Text Processing	53
3.4.2	Intrinsic Evaluation	54
3.4.3	Comparative Evaluation	54
3.5	Conclusions	57

4.1 Introduction

Including external information in the text analysis tasks have received much attention in recent years. Particularly, the use of publicly available resources is one of the most trending topics in the research community. One of the most interesting cases is the introduction of Wikipedia information in unsupervised or supervised learning [Scaiella et al., 2012, Milne and Witten, 2013, Lau et al., 2013]. For instance, current open NLP problems such as word sense disambiguation, semantic similarity and conference resolution - to name a few - have been successfully addressed using techniques based on external information. In particular, the use of Wikipedia information in the SRC problem has been studied in [Scaiella et al., 2012].

As discussed before, most successful methodologies follow a monothetic approach with some exceptions that follow the polythetic approach. In this Chapter, our research is motivated by the fact that the adequate combination of the polythetic and monothetic approaches in a single algorithm should lead to improved performance by three important factors in SRC: *clustering accuracy*, *labeling quality* and *partitioning shape*. In Section 3.2, we presented the SRC algorithm capable of dealing with a polythetic solution. However, as mentioned in the conclusions previously, excessive use of words in cluster labels could introduce additional effort in users' label understanding. Because of this, we present a new algorithm called *Dual C-Means*, which provides a theoretical background for dual-representation clustering. Its originality relies on the fact that different representation spaces can drive the clustering process as well as the labeling task in a single step. In particular, we propose to introduce external information to improve cluster label quality and analyze their impact in the clustering performance. As external information we refer to information obtained from a source out of the Web snippets. Traditional SRC systems obtain the cluster first and then assign labels to it or vice versa. However, label quality or cluster quality is not guaranteed. To overcome this situation, the external information is a source of well-formed labels such as query logs, anchor text, predefined categories, etc. that help in the labeling process but that also drive the clustering process.

In terms of evaluation, we test the proposed algorithm with different metrics (i.e., F_1^N [Di Marco and Navigli, 2013], F_{b^3} [Amigó et al., 2009], ARI [Vinh et al., 2009], $D\#-nDCG$ [Sakai and Song, 2011]), over well-studied datasets (e.g. ODP-239 [Carpineto and Romano, 2010], SEMEVAL [Navigli and Vannella, 2013]) and different representation spaces (i.e. text and query logs). The results show that the combination of the polythetic representation of Web snippets and a query log based representation of cluster centroids achieve the best configuration for the SRC task. In particular, increased performance is shown against most SRC solutions (i.e. STC [Zamir and Etzioni, 1998], LINGO [Osinski and Weiss, 2005], TOPICAL [Scaiella et al., 2012], LDA [Blei et al., 2003]). Our main contributions in this chapter are:

- A new algorithm (Dual *C-Means*), which can be seen as an extension of *K-means* [Lloyd, 1982] for dual-representation spaces;
- An instantiation of the Dual *C-Means* for SRC, which takes advantage of external resources such as improving clustering accuracy, labeling quality and partitioning shape;
- A new annotated dataset (WEBSRC401) based on the TREC Web Track 2012 for full SRC evaluation over the Web.

4.2 Dual C-means Algorithm

This section is devoted to the presentation of the Dual C -Means algorithm that extends the classical K -means [Lloyd, 1982] for dual representation spaces. In the first subsection, we present the general model and in the second, we propose its instantiation for the specific task of SRC.

4.2.1 General Model

Let S be a dataset to partition where each data $s_i \in S$ is described on a representation space E_1 and E_2 denotes another space supporting cluster representation. We hypothesize the existence of a function $d : E_1 \times E_2 \rightarrow \mathbb{R}^+$ quantifying the dissimilarity between any data from E_1 and any cluster representative from E_2 . The newly proposed clustering model (Dual C -Means) is driven by the objective criterion defined in Equation 4.1, which must be minimized.

$$J_{dcm}(\Pi, M) = \sum_{k=1}^c \sum_{s_i \in \pi_k} d(s_i, m_k) \quad (4.1)$$

As illustrated in Figure 4.1, the aim of the minimization of $J_{dcm}(\Pi, M)$ is to find a partition of S into c clusters ($\Pi = \{\pi_1, \dots, \pi_c\}$) so that in each cluster π_k , any object is as closed as possible to a common cluster representative m_k ($M = \{m_1, \dots, m_c\}$).

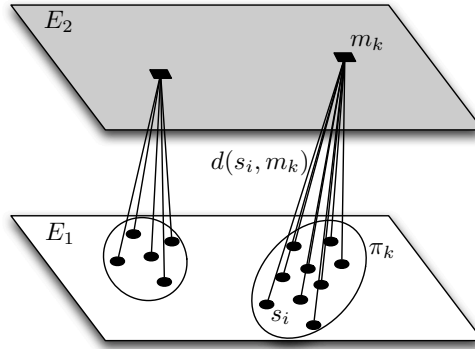


Figure 4.1: Dual C -Means aims to discover clusters of objects in E_1 closed to a common cluster representative in E_2 .

The optimization process can be achieved by an usual dynamic reallocation algorithm starting with a random initial clustering Π^0 and then iterating the following two steps (Update and Assignment) until convergence:

- (i) Update: compute new optimal cluster representatives M^{t+1} considering a fixed partition Π^t ,

- (ii) Assignment: compute new optimal assignments Π^{t+1} considering fixed cluster representatives M^{t+1} and use the following rule to assign each object to its closest representative:

$$\forall s_i, s_i \in \pi_k \Leftrightarrow k = \arg \min_{l=1, \dots, c} d(s_i, m_l).$$

Note that the updating of cluster representatives has to be defined according to both the dissimilarity measure $d(.,.)$ and the representative space E_2 in order to ensure that the objective criterion $J_{dcm}(.,.)$ decreases. Let us also notice that in the specific case where $E_1 = E_2 = \mathbb{R}^n$ and the squared euclidean distance is chosen as dissimilarity $d(.,.)$, the Dual C -Means algorithm comes down exactly to the usual K -means algorithm ($m_k^{t+1} = \sum_{s_i \in \pi_k^t} s_i / |\pi_k^t|$). Finally, as K -means, Dual C -Means is sensitive to random initialization and requires the number of expected clusters (C) as parameter¹.

4.2.2 Instantiation in the SRC Context

In SRC objects are normally Web snippets represented in the E_1 space ($s_i \in S$) and cluster representatives are labels represented in the E_2 space ($m_k \in M$).

The crucial hypothesis of the Dual C -Means algorithm is the existence of a dissimilarity metric $d(.,.)$ capable of comparing objects from different feature spaces. For this reason, a matching process of the two feature sets is required that can be formalized as a transition matrix P ($p_1 \times p_2$) quantifying this matching for each of the p_1 features defined in E_1 with each of the p_2 features from E_2 .

Without loss of generality, we define a generic dissimilarity measure, considering a transition matrix as in Equation 4.2 where m_k^T is the transposed label vector, $s_i P m_k^T$ quantifies a similarity between a Web snippet s_i and a label m_k , and α is a constant to adjust in order to ensure dissimilarity values in \mathbb{R}^+ .

$$d(s_i, m_k) = \alpha - s_i P m_k^T \quad (4.2)$$

A dissimilarity form such as this allows us to rewrite the Dual C -Means algorithm as a maximization problem defined in Equation 4.3.

$$\min_{\Pi, M} \sum_{k=1}^c \sum_{s_i \in \pi_k} d(s_i, m_k) \Leftrightarrow \max_{\Pi, M} \sum_{k=1}^c \sum_{s_i \in \pi_k} s_i P m_k^T \quad (4.3)$$

Notice that when the label space E_2 is unconstrained (e.g. $E_2 = \mathbb{R}^{p_2}$), the resolution of Equation 4.3 has no sense ($M = +\infty$). But in SRC, a small set of words (i.e. the labels) are usually chosen to help the user in his search for information. Thus,

¹This issue will be tackled in Section 4.4.

we consider two vocabularies V_1 and V_2 defining the two feature spaces E_1 and E_2 respectively. We constrain Web snippet descriptions to be word distributions over V_1 ($s_{i,j} \in [0, 1] \forall i, j$ and $\sum_{j=1}^{p_1} s_{i,j} = 1$) and cluster labels to subsets of p words from V_2 ($E_2 = \{m_k \in \{0, 1\}^{p_2} \mid \sum_{l=1}^{p_2} m_{k,l} = p\}$).

In the above case, the computation of optimal cluster labels is a discrete optimization process solved for each cluster π_k independently, first by sorting the vocabulary V_2 from the most representative word (l_1^k) to the least representative one ($l_{p_2}^k$) using the representativity function defined in Equation 4.4

$$\forall l, k \quad \lambda_k(l) = \sum_{s_i \in \pi_k} s_i P_{.,l} \quad (4.4)$$

and then by defining a cluster label vector m_k as the combination of the p most representative words from V_2 for the snippets in π_k as proposed in Equation 4.5.

$$m_{k,l} = \begin{cases} 1 & \text{if } l \geq l_p^k \\ 0 & \text{otherwise} \end{cases} \quad (4.5)$$

It is good to notice that the SRC algorithm proposed in Section 3.2 falls into an SRC instantiation of the Dual C-Means algorithm if the following constraints are true:

- Web snippet and label representation spaces are not dissociated (i.e. $V_1 = V_2$) thus not benefiting from the duality of the clustering algorithm;
- The transition matrix P is computed with the Symmetric Conditional Probability (SCP [Silva et al., 1999]) or the Pointwise Mutual Information (PMI [Church and Hanks, 1990]) with the unique vocabulary $V_1 = V_2$.

So, the Dual C-Means algorithm can be considered as a generalization of SRC algorithms. In particular, the advances - with respect to the automatic definition of cluster size - can be directly applied in the Dual C-Means algorithm. The study of this characteristic is discussed in Section 3.3².

To make use of the duality concept of the newly proposed algorithm in SRC, we would suggest differentiating the two vocabularies V_1 and V_2 . First, V_1 is defined as the bag of words occurring in all Web snippets retrieved for a given query. Second, if we consider a set Y of any kind of consistent external information, such as query logs, anchor texts, predefined categories, etc., the vocabulary V_2 is defined as the bag of words occurring in Y . E_2 is restricted to the set of external candidates³ defined as distributions in the vector

²In this Chapter, we concentrate our efforts in the evaluation results when the presented instantiation is used.

³Those obtained from the external information source.

space model induced by V_2 . This situation is formalized in Equation 4.6 with β_i denoting the size of the external candidate y_i .

$$E_2 = \{y_i \in \{0, \frac{1}{\beta_i}\}^{p_2} \mid \sum_{j=1}^{p_2} y_{i,j} = 1 \text{ and } y_i \in Y\} \quad (4.6)$$

This kind of clustering is polythetic but driven by the labels proposed by the external information. Figure 4.2 illustrates the instantiation of the Dual C -Means algorithm within SRC, where the restricted set of available external candidates guides the cluster formation process.

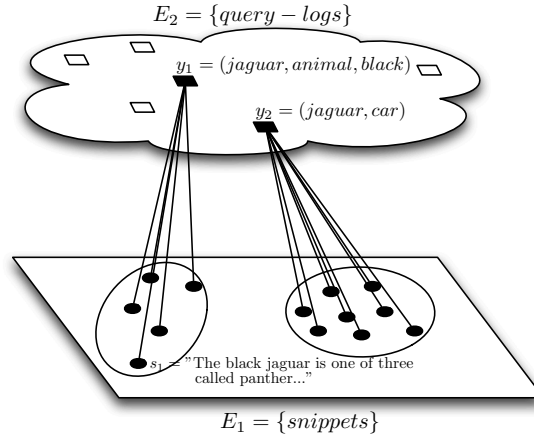


Figure 4.2: Example of the Dual C -Means instantiated for the SRC context with external information labels as cluster label space.

It is important to realize that the Dual C -Means algorithm does not restrict the source to generate the E_2 space to a specific kind of external information. Indeed, any of the previously mentioned source could cast in the presented instantiation. However, for evaluation purposes, we selected query logs for the external candidates, and as a consequence, query logs databases for the external information used for Dual C -Means. Indeed, as shown in Section 2.2.3 in the user evaluation, query logs are meaningful labels.

It is important to note that any actual SRC dataset does not include an automatic way to perform label evaluation⁴ with query logs. For this reason, we have developed a dataset for the SRC problem that support query logs for labeling evaluation.

⁴Note that kSSL can be automatically calculated, but relevance labels must be provided for each subtopic and each evaluated SRC algorithm.

4.3 The WEBSRC401 Dataset

Most commonly used datasets in SRC are described in Section 2.3.2. Although ODP-239 and SEMEVAL are the most reliable SRC datasets⁵, they do not suit when a unique framework for evaluate clustering quality and labeling quality is desired.

ODP-239 does not represent the typical kind of results obtained by querying using any given search engine as the number of possible subtopics is always equal to 10. It is obvious that this structure clearly differs from a typical Web results set. Moreover, queries are not extracted from query logs but rather chosen based on the categories present in DMOZ.

On the other hand, the subtopics in SEMEVAL follow a more natural distribution as they are defined based on the disambiguation pages of Wikipedia. As such, these subtopics are likely to cover most of the senses present in the Web for the 100 evaluated queries. However, SEMEVAL is built to specifically deal with ambiguous queries, which are self contained in Wikipedia. It is, however, clear that not all queries are Wikipedia articles or ambiguous. For example, many queries are multifaceted but not ambiguous [Kong and Allan, 2013]. Take “Olympic Games” for example. Its Wikipedia entry is not ambiguous but it represents many different facets such as history, logos, dates or cities, to name just a few.

Consequently, it is clear that different results will be obtained from one dataset to another⁶. To afford a more realistic situation in Web search results, we propose a new SRC dataset based on the ClueWeb09 Category B text collection (CCB)⁷, which comprises about 50 million pages in English, including the entirety of the English Wikipedia and the task descriptions of the TREC Web Track 2012. The goal of TREC Web Track 2012 is to return a ranked list of Web pages that together provide complete topical coverage of any given query, while avoiding excessive redundancy of the subtopics in the result list.

In particular, each topic contains a query field, a description field and several subtopic fields which can be ambiguous or multifaceted. For each topic, a judgment file (i.e. qrel⁸) includes the list of relevant Web pages from CCB and the manually attributed grade of the Web page subtopic.

Instead of retrieving relevant Web pages, we are interested in obtaining relevant clusters (i.e. Web pages with the same subtopic) that have high coverage of all the subtopics. So, we propose to transform the data available in the TREC Web Track 2012 in a typical SRC

⁵Remember that Moresque does not fulfil classical SRC dataset properties. See Section 3.4.3.

⁶A quick summary of both datasets is presented in Table 2.3.

⁷<http://lemurproject.org/clueweb09/> [Last access: 27/01/2014]

⁸The qrel files are text files with the document names and a numeric values which indicates their relevance. Usually, zero is used to indicate that is not a relevant document and positive integer values are used for relevant ones.

format [Carpineto and Romano, 2010], the result of which is the WEBSRC401 dataset⁹. First, for each Web page considered as query-relevant, its Web snippet is retrieved using the *SnippetGenerator* function of ChatNoir¹⁰. By default, a Web snippet composed of a maximum of 500 characters found around the query words is provided. Examples of queries, subtopics and Web snippets can be found in Table 4.1.

Second, for each query, its subtopics are defined as in the TREC Web Track 2012 and each qrel is encoded in a new format, which contains the Web page id, the subtopic id and the query¹¹. Additionally, it is important to notice that the WEBSRC401 dataset facilitates the evaluation of new techniques based on more complex resources provided by researchers as it is based on the well-studied ClueWeb09. For example, cluster ranking or spam cluster filtering studies could be endeavored in with the PageRank scores and the spam rankings of ClueWeb09 which are publicly available.

4.4 Clustering Evaluation

As mentioned in [Carpineto et al., 2009], evaluating SRC systems is a hard task. Indeed, the evaluation process is three-fold. A successful SRC system must discover relevant topical clusters (*clustering accuracy*) and propose meaningful labels at the same time (*labeling quality*). We will also see in our experiments that *partition shape* is also an important factor to study.

4.4.1 Evaluation of SRC

First, a successful SRC system must evidence high quality level clustering. Ideally, each query subtopic should be represented by a unique cluster containing all the relevant Web pages inside. However, this task is far from being achievable so far. Therefore, this constraint can be reformulated as for the TREC Web Track 2012: the task of SRC systems is to provide complete topical cluster coverage of any given query, while avoiding excessive redundancy of the subtopics in the result list of clusters.

Secondly, SRC systems should present meaningful labels to the user to ease his search for information. Therefore, the evaluation of the labeling task is of utmost importance. As far as we know, only [Carpineto and Romano, 2010, Scaiella et al., 2012] propose evaluating both dimensions. However, their experiments are not reproducible as they rely on manually annotated datasets¹², which are not publicly available.

⁹<https://websrc401.greyc.fr/> [Last access: 27/01/2014].

¹⁰<http://chatnoir.webis.de/> [Last access: 27/01/2014].

¹¹Note that these steps could be used to extend the dataset with the TREC Web tracks of the years 2009, 2010 and 2011.

¹²ODP-239 is publicly available. However, the annotations to calculate k SSL are not available.

Query	Subtopics	Examples (Url/Web snippets)
the beatles rock band	Find a listing of Beatles songs.	http://artists.letssingit.com/the-beatles-qfjds/biography
		British rock guitarist/singer based in Hamburg, Tony Sheridan. The Beatles hadn't fully developed at this point, and these recordings many of which...
	What is the history of the Beatles rock band?	http://classicrock.about.com/od/beatles/a/beatles_history.htm
		Lennon's schoolboy skiffle group, The Quarry Men, formed in 1957, eventually became The ... group, The Quarry Men, formed in 1957, eventually became...
	What albums did the Beatles release?	http://home.att.net/~chuckayoub/the_beatles_lyrics.html
		The Beatles were also the first British rock group to achieve worldwide prominence, launching a British Invasion that made rock truly an international phenomenon...
	What are the names of the members of the Beatles?	http://www.rollingstone.com/rockdaily/index.php/2007/06/27/larry-king-calls-ringo-george-on-live-tv/
		Yesterday we mentioned we were worried how Larry King would fare interviewing Paul McCartney , Ringo Starr , Yoko Ono and Olivia Harrison. We accurately guessed that King would...
grilling	Find kabob recipes.	http://bbq.about.com/od/grillinghelp/a/aa073005a.htm
		Take your favorite cocktail and turn it into a great grilled dish Grilling Help For centuries people have used wine and beer as marinades...
	Find tips on grilling vegetables.	http://allrecipes.com/HowTo/Grilling-101-Grilled-Vegetables/Detail.aspx
		Summer grilling often conjures images of testosterone-addled men wrestling slabs of meat, but let's consider another eminently grillable foodstuff--the vegetable. Great...
	Find tips on grilling fish.	http://allrecipes.com/HowTo/Plank-Grilling/Detail.aspx
		Fish is the original favorite, but grilling with wood planks will introduce a whole new range of savory flavors to veggies, meats and more...
	Find instructions for grilling chicken.	http://bbq.about.com/od/grillinghelp/u/grilling.htm
		Guide to Barbecues & Grilling If it can be roasted, broiled, sautd, fried or baked then it can probably be grilled...
	Find the Grilling Magazine website.	http://www.grillingmag.com/
		Get the latest Grilling and Barbecue Recipes at our online Grilling recipes center. Learn new Grilling techniques from Grilling Chefs worldwide...

Table 4.1: Text query and associated subtopics for the queries ids 155 and 160 in the WEB-SRC401 dataset. The Web snippets were selected from the list of relevant documents that were manually annotated in the TREC Web Track 2012.

In the following sections, we propose a complete set of repeatable experiments to give an exhaustive overview of the SRC field. We start by focusing on the experimental setups.

4.4.2 Experimental Setups

In this section, we compare different configurations of the Dual C -Means to different state-of-the-art algorithms using well-studied evaluation metrics.

Dual C -Means Configurations The originality of the Dual C -Means is its embodiment of a great number of possible configurations due to the expressiveness of its model. In this section, we will particularly focus on two main issues. The first one deals with using different similarity measures to compute the transition matrix P . The general idea is supported by the fact that different word similarity measures produce different results [Pecina and Schlesinger, 2006]. Consequently, we aim to understand their impact on the SRC task. The second one aims to test our initial hypothesis by stating that the introduction of external resources can improve SRC. As a consequence, we propose two different space representations: text-text (i.e. $V_1 = V_2$) and text-query logs (i.e. $V_1 \neq V_2$).

Word Similarity Measures The use of word similarity measures is an important and replaceable component of our algorithm encoded in the transition matrix. In this study, we compare a total of five collocation measures¹³. We used¹⁴ the Symmetric Conditional Probability (SCP) [Silva et al., 1999] in Equation 4.7, the Pointwise Mutual Information (PMI) [Church and Hanks, 1990] in Equation 4.8, the Dice coefficient [Dice, 1945] in Equation 4.9, the LogLikelihood ratio (LogLike) [Dunning, 1993] in Equation 4.10 and Φ^2 [Gale and Church, 1991] in Equation 4.11. The expressiveness of the Dual C -means allows the definition of different types of word similarity measures. As a consequence, we also compute word-word similarity based on the VSM representation. Basically, for each snippet $s_i \in S$, a simple word-word similarity measure is $S^T S$ where S^T is the transposed of the snippet-term matrix S . In this case, $P = S^T S$. Another interesting similarity measure is LSA [Landauer and Dumais, 1997], which can be formulated as follows: $P = U \Lambda_e U^T$ where $U \Lambda U^T$ is the eigen decomposition of $S^T S$, and e is the number of highest eigen values selected to represent the latent space¹⁵.

$$SCP(w_i, w_j) = \frac{P(w_i, w_j)^2}{P(w_i) * P(w_j)} \quad (4.7)$$

¹³It is clear that there exist a great deal of association measures that could be tested or other methods for word similarity calculation. However, we selected the ones, which best complement themselves.

¹⁴Note that SCP and PMI are already defined in Section 3.2.3, but we have included again here for simplicity.

¹⁵In our experiments, this value was set to the minimum which guarantees that $\sum_{i=1}^e \Lambda_i \geq 0.9 \sum_{i=1}^{p1} \Lambda_i$.

$$PMI(w_i, w_j) = \log_2 \frac{P(w_i, w_j)}{P(w_i) * P(w_j)} \quad (4.8)$$

$$DICE(w_i, w_j) = \frac{2 * f(w_i, w_j)}{f(w_i) + f(w_j)} \quad (4.9)$$

$$\begin{aligned} LogLike(w_i, w_j) = & -2 * logLike(f(w_i, w_j), f(w_i), \frac{f(w_j)}{N}) \\ & + logLike(f(w_j) - f(w_i, w_j), N - f(w_i), \frac{f(w_j)}{N}) \\ & - logLike(f(w_i, w_j), f(w_i), \frac{f(w_i, w_j)}{f(w_i)}) \\ & - logLike(f(w_j) - f(w_i, w_j), N - f(w_i), \frac{f(w_j) - f(w_i, w_j)}{N - f(w_i)}) \end{aligned} \quad (4.10)$$

$$\Phi^2(w_i, w_j) = \frac{P(w_i, w_j) - P(w_i) * P(w_j)}{P(w_i) * P(w_j) * (1 - P(w_i)) * (1 - P(w_j))} \quad (4.11)$$

where $logLike(a, b, c) = (a * Log(c)) + ((b - a) * Log(1 - c))$, $P(w_i, w_j)$ is the joint probability of the words w_i and w_j , $P(w_i)$ is the marginal probability of the word w_i , $f(w_i, w_j)$ is the frequency of word pairs (w_i, w_j) , $f(w_i)$ is the frequency of the word w_i and N is the number of retrieved Web snippets.

SRC Algorithms We aim to compare our algorithm to the most competitive strategies proposed so far in SRC literature. For that purpose, we show the results of TOPICAL [Scaiella et al., 2012], LDA [Blei et al., 2003]¹⁶ and all other strategies mentioned in previous chapters¹⁷, i.e., STC [Zamir and Etzioni, 1998] and LINGO [Osinski and Weiss, 2005]. It is worth noticing that for evaluation purposes, we have developed an open source implementation¹⁸ of TOPICAL using the Wikipedia Miner API proposed by [Milne and Witten, 2013] and the spectral algorithm proposed by [Ng et al., 2001] included in the SCIKIT learning tool¹⁹. And for LDA, we used the topic modeling package included in the MALLET toolkit [McCallum, 2002]. The parameters were set following the toolkit instructions (i.e. stop-words removal, $\alpha_t = 0.01$, $\beta_w = 0.01$ and limited to 1000 iterations) and the cluster membership was assigned by taking the maximum topic probability value. Other algorithms were used similarly as mentioned in

¹⁶It is clear that LDA is not a SRC algorithm, but it is included for comparison with classical text clustering algorithms, as it has recently been proposed in the semeval competition [Lau et al., 2013].

¹⁷Values for AGK-means are not included because Dual C-means is a more general algorithm that obtains similar results when equivalent configurations are used.

¹⁸This implementation is publicly available upon request to the authors.

¹⁹<http://scikit-learn.org/stable/> [Last access: 27/01/2014].

Chapters 2 and 3.

Evaluation Metrics The full definition of the evaluation metrics can be found in Section 2.3.2. For the implementation, we used the Java evaluator²⁰ to compute the F_1^N evaluation metric, and the implementation provided by [Amigó et al., 2013]²¹ to compute F_{b3} . Finally, for F_1^C , we use our implementation of this metric²².

Text Processing and Implementation To provide a more reproducible framework, we replaced the preprocessing step defined in 3.4.1 with a more classic one. For this reason, all Web snippets were tokenized with the GATE platform²³ but we did not apply the removal of stop-words. For the dynamic reallocation algorithm, we used the optimized version of K -means++ proposed in [Arthur and Vassilvitskii, 2007] as the initialization process is semi-deterministic²⁴ and there exists an efficient implementation called Scalable K -means++ [Bahmani et al., 2012].

4.4.3 Clustering Results

A great deal of experiments have been performed to achieve conclusive results. We first propose to evaluate clustering accuracy of the Dual C -Means against different state-of-the-art algorithms. For this purpose, we propose an exhaustive search as in [Scaiella et al., 2012], the underlying idea of which is to evaluate the behavior of a given algorithm along with the increasing number of output partitions. In this first set of experiments, we pretend to understand the clustering quality of our approach when only text information is taken into account (i.e. $V_1 = V_2$ and the number of p words composing the centroids is set to 2^{25}) and compare it to state-of-the-art algorithms. We present the results for 19 runs ($K = 2..20$) and illustrate the F_{b3} values over ODP-239 and WEBSRC401 in Figure 4.3 and 4.4.

The obtained results show interesting situations. In all the cases, Dual C -means outperforms state-of-the-art algorithms in terms of clustering accuracy. In particular, SCP, DICE and LogLike show improved results and outperform other word-word similarity metrics. It is interesting to notice that PMI and Φ^2 - which are known to give less importance to more frequent events - show less relevant results.

²⁰<http://www.cs.york.ac.uk/semeval-2013/task11/index.php?id=data> [Last access: 27/01/2014].

²¹<http://nlp.uned.es/~enrique/software/RS.zip> [Last access: 27/01/2014].

²²<https://sites.google.com/site/jgmorenof/josemoreno/easy-web-search-results-clustering> [Last access: 27/06/2014].

²³<http://gate.ac.uk/> [Last access: 27/01/2014].

²⁴Note that for our experiments, the first seed Web snippet is selected as the one that is most similar to all the other ones in S .

²⁵Note that the selection of $p = 2$ is not the best configuration, but it is enough to achieve good performance, and is better tuned for SRC as processing power is limited.

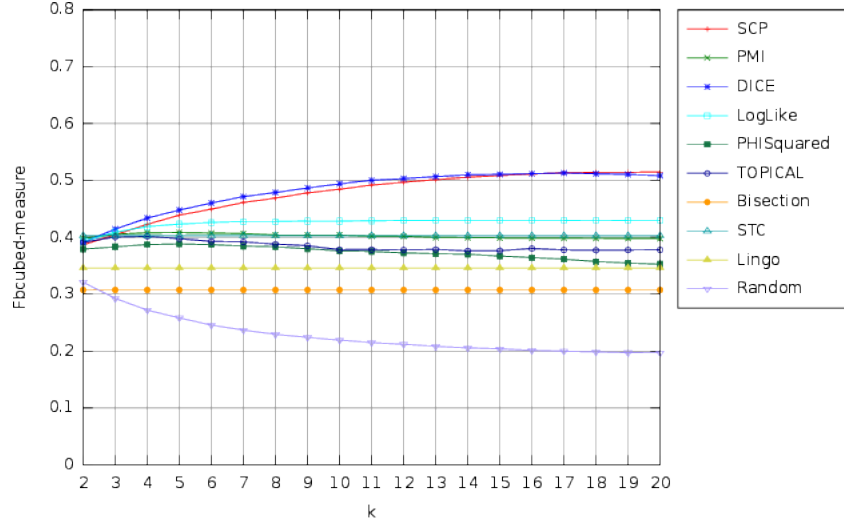


Figure 4.3: Impact of K for F_{b3} against ODP-239 dataset.

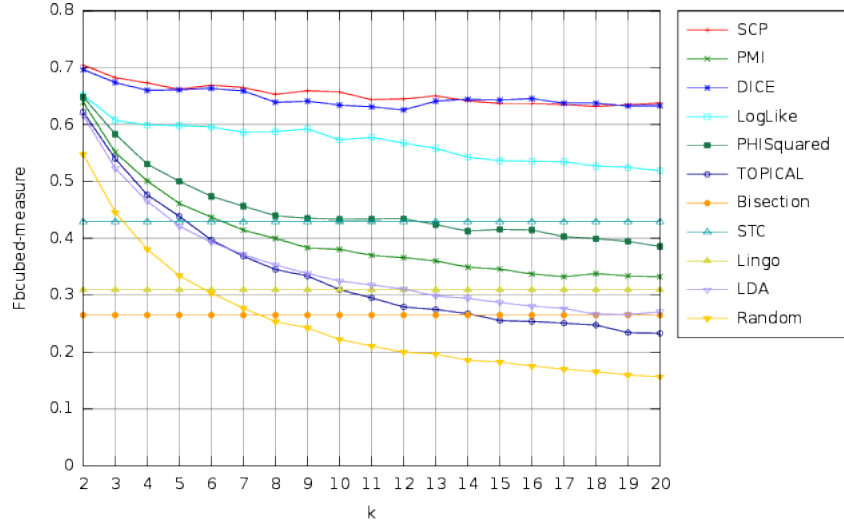


Figure 4.4: Impact of K for F_{b3} against WEBSRC401 dataset.

These results only give a small idea of the overall phenomena. In Tables 4.2, 4.3 and 4.4, results for 10 cluster outputs are given for all metrics and all datasets. These new results show interesting properties of evaluation metrics. Although Dual C -means shows improvements over all the other competitors in terms of F_{b3} or F_1^C (except in one case) for ODP-239, SEMEVAL and WEBSRC401, this situation does not stand for the other metric, F_1^N . For ODP-239, the best results are obtained LINGO in terms of F_1^N . For SEMEVAL, the best performances are provided by LINGO in terms of F_1^N . On the other hand, F_1^N shows inconsistent results when compared to all the other metrics. In particular, it tends to give high results when the other metrics decrease.

Although different results are obtained for SEMEVAL and ODP-239, steady results are obtained for WEBSRC401 by the Dual C -Means configured with the $S^T S$ word-word similarity metric. Indeed, it clearly outperforms all other algorithms in terms of F_{b3} and

F_1^C . At this stage in our experiments, we can conclude that this configuration provides the best performance - both in terms of clustering accuracy and partitioning shape.

		F_1^N	F_{b3}	F_1^C
ODP-239	TOPICAL	0.5760	0.3799	0.2839
	LDA	0.5978	0.4370	0.3900
	STC	0.5499	0.4027	0.3238
	LINGO	0.6636	0.3461	0.2029
SEMEVAL	TOPICAL	0.6791	0.3998	0.2723
	LDA	0.7159	0.3966	0.2840
	STC	0.7223	0.4632	0.3682
	LINGO	0.7742	0.3662	0.2072
WEBSRC401	TOPICAL	0.6932	0.3083	0.2522
	LDA	0.7020	0.3214	0.2613
	STC	0.6779	0.4293	0.3905
	LINGO	0.7123	0.3095	0.2502

Table 4.2: Results of the state-of-the-art algorithms for ODP-239, SEMEVAL and WEBSRC401. K was fixed to 10 clusters for TOPICAL and LDA. Other algorithms were ran using standard configurations.

		F_1^N	F_{b3}	F_1^C
ODP-239	SCP	0.4961	0.4845	0.3785
	PMI	0.5671	0.4041	0.3231
	DICE	0.5181	0.4939	0.3885
	LOGLIKE	0.5078	0.4285	0.3650
	Φ^2	0.5479	0.3759	0.3059
	$S^T S$	0.5294	0.4852	0.3822
	LSA	0.5482	0.4712	0.3731
SEMEVAL	SCP	0.6114	0.5632	0.4856
	PMI	0.6634	0.4198	0.3297
	DICE	0.6245	0.5763	0.4914
	LOGLIKE	0.5753	0.5416	0.4934
	Φ^2	0.6797	0.3972	0.2932
	$S^T S$	0.6225	0.5722	0.4808
	LSA	0.6219	0.5645	0.4684

Table 4.3: Results of the Dual C -Means algorithm for ODP-239 and SEMEVAL. K fixed to 10 Clusters. Let us notice that for all experiments, the number of p words composing the centroids was set to 2 and the vocabulary is the set of words appearing in the retrieved Web snippets.

The second set of our experiments aims to analyse the behavior of Dual C -Means when external resources are included. In this case, we use the set of query logs provided by the NTCIR-10 Intent-2 task [Sakai et al., 2013a] and propose to drive the clustering process by this external information. Therefore, a cluster centroid is represented by its most representative query log. Results are presented in Table 4.4 where $V_1 \neq V_2$ for WEBSRC401. Let us notice that this is the only dataset for which experiments with query logs can be

		F_1^N	F_{b3}	F_1^C
$V_1 = V_2$ (Text)	SCP	0.6698	0.6597	0.6217
	PMI	0.6788	0.3981	0.3514
	DICE	0.6718 †	0.6575	0.6202
	LOGLIKE	0.6566	0.5499	0.5131
	Φ^2	0.6841	0.4299	0.3836
	$S^T S$	0.6713	0.6666 †	0.6260 †
	LSA	0.6706	0.6327 †	0.5884 †
$V_1 \neq V_2$ (QL)	SCP	0.6580	0.6572	0.6239
	PMI	0.6866	0.3806	0.3338
	DICE	0.6593	0.6343	0.6023
	LOGLIKE	0.6636	0.5728	0.5394
	Φ^2	0.6783	0.4333	0.3926
	$S^T S$	0.6645	0.6160	0.5847
	LSA	0.6719	0.5577	0.5264

Table 4.4: Results of the Dual C-Means algorithm for WEBSRC401. K was fixed to 10 clusters. Note that for the experiments when $V_1 = V_2$, the number of p words composing the centroids was set to 2 and the vocabulary is the set of words appearing in the retrieved Web snippets. Note that † means paired student's t -test statistical relevance for p - value < 0.05 between a given metric in $V_1 = V_2$ and its counterpart in $V_1 \neq V_2$.

performed and easily reproduced.

Not surprisingly, the introduction of external information decreases clustering accuracy. This situation is due to the limited possibilities when the query logs are used. Indeed, when $V_1 \neq V_2$ possible centroids are any combination of p words and when $V_1 = V_2$ the centroids candidates come from the limited query logs database. The cluster accuracy decreasing is true only at a glimpse when comparing $S^T S$ for $V_1 = V_2$ and SCP for $V_1 \neq V_2$ (statistical relevance is not true in this case). The most relevant benefit of this new dual approach is embodied by the expressiveness of the query logs as meaningful labels. The evaluation of this benefit is the objective of the next section.

4.5 Labeling Evaluation

As mentioned in [Carpineto et al., 2009], the labeling process plays an important role in the success of SRC systems. As a consequence, a clear objective evaluation is needed. However, this has not yet been the case. Indeed, [Hearst and Pedersen, 1996][Ferragina and Gulli, 2008] proposed user studies, which are difficult to replicate. In order to solve this problem, [Carpineto and Romano, 2010][Scaiella et al., 2012] proposed to evaluate the k SSL metric but their datasets are defined in two different ways and they are not publicly available. So, in order to propose a conclusive evaluation of the labeling process, we propose to use a new gold standard dataset provided by the Subtopic Mining subtask

of the NTCIR-10 Intent-2 [Sakai et al., 2013a] and apply recent evaluation metrics proposed by [Sakai and Song, 2011]: $I-rec@10$, $D-nDCG@10$ and $D\#-nDCG@10$. These metrics aim to measure Precision and Recall of the users' intents.

In SRC, we can use the labels provided by the algorithms as the users' intents candidates. If so, we can directly apply the given metrics. So then, $I-rec$ measures the number of intents discovered by the algorithm over the total different intents of the query. This metric can simply be viewed as an intent Recall. Then, $D-nDCG$ is obtained by sorting all relevant intents by the global gain, which is defined as the sum of all the individual intent gains. Finally, the $D\#-nDCG$ metric is the linear combination of $I-rec$ and $D-nDCG$, using γ and $1 - \gamma$ factors. Note that defining the probabilities of each intent as well as the relevant intents can be a hard task. However, as our experiments are realized over WEBSRC401 based on ClueWeb09, these values are known and publicly available [Sakai et al., 2013a]. The NTCIREVAL toolkit²⁶ was used for the calculation of these metrics. Let us notice that for the specific task of SRC, we propose to use $I-rec@10$, $D-nDCG@10$ and $D\#-nDCG@10$ as for most queries the number of intents is limited. These metrics are defined in the Equations 4.12, 4.13 and 4.14.

$$I-rec@N = \frac{|I'|}{|I|} \quad (4.12)$$

where I is the set of known intents for a query q and I' is the set of intents covered by the returned labels at level N .

$$D-nDCG@N = \frac{\sum_{r=1}^N \sum_i Pr(i|q) g_i(r) / \log(r+1)}{\sum_{r=1}^N \sum_i Pr^*(i|q) g_i^*(r) / \log(r+1)} \quad (4.13)$$

where $Pr(i|q)$ (resp. $Pr^*(i|q)$) denotes the intent probability obtained for the discovered labels (resp. for the reference labels) and $g_i(r)$ (resp. $g_i^*(r)$) is the gain value of the label at rank r with respect to i for the output of the labeling (resp. for the reference labeling).

$$D\#-nDCG@N = \gamma I-rec@N + (1 - \gamma) D-nDCG@N \quad (4.14)$$

where γ was set to 0.5 following the framework evaluation proposed in the Subtopic Mining subtask of the NTCIR-10 Intent-2.

The results provided by [Sakai et al., 2013b] for different query completions ($Bing_C$, $Google_C$ and $Yahoo_C$), query suggestions ($Bing_S$) services and a simple merging strategy (Merge) are reported in Table 4.5 as well as the results of our approach. In particular, we show the results when clustering is query log driven ($V_1 \neq V_2$) and when labeling is performed *a posteriori* ($V_1 = V_2$). By *a posteriori*, we mean that clustering is first performed

²⁶<http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html> [Last access: 27/01/2014].

on the exclusive text representation. Then, for the second step, the label is computed by any heuristic. In our experiments, the query log that best represents each text-based cluster is computed using one iteration of the update function defined in section 4.2, which allows direct comparison of the results.

		$I - rec@10$	$nDCG@10$	$D\# - nDCG@10$
$V_1 = V_2$	SCP	0.2804	0.3195	0.2959
	PMI	0.3136	0.3444	0.3250
	DICE	0.2952	0.3242	0.3093
	LOGLIKE	0.2269	0.2885	0.2550
	Φ^2	0.3390	0.3642	0.3523
	$S^T S$	0.2837	0.3063	0.2935
	LSA	0.3238	0.3694	0.3456
$V_1 \neq V_2$	SCP	0.3669 †	0.3932 †	0.3793 †
	PMI	0.4136 †	0.4257 †	0.4203 †
	DICE	0.3761 †	0.3884 †	0.3814 †
	LOGLIKE	0.3937 †	0.4146 †	0.4046 †
	Φ^2	0.4249 †	0.4221 †	0.4225 †
	$S^T S$	0.4033 †	0.4273 †	0.4119 †
	LSA	0.3946 †	0.4197 †	0.4050 †
Baselines	BingS	0.3068	0.2787	0.2928
	BingC	0.3231	0.3268	0.3250
	GoogleC	0.3735	0.3841	0.3788
	YahooC	0.3829	0.3815	0.3822
	Merge	0.3365	0.3181	0.3273

Table 4.5: Evaluation results of the labeling process with query logs over the NTCIR-10 Intent-2 dataset. Note that † means paired student’s t-test statistical relevance for $p - value < 0.05$ between a given metric in $V_1 = V_2$ and its counterpart in $V_1 \neq V_2$.

The results of the query driven Dual C -Means outperform all baselines and a *posteriori* labeling. Moreover, all the differences between a given metric in $V_1 = V_2$ and its counterpart in $V_1 \neq V_2$ are statistically relevant. These results also show interesting behaviors. Indeed, while PMI and Φ^2 collocation metrics previously showed worst clustering accuracy results compared to other configurations, they show improved results in terms of labeling. The fact that these metrics tend to favour less frequent associations is an interesting characteristic for labeling purposes and a conclusive remark. Moreover, the $S^T S$ word-word similarity measure shows high $nDCG@10$ value and competitive overall $D\# - nDCG@10$. These results clearly point towards this last configuration as the best compromise for clustering accuracy, labeling quality and partitioning shape.

4.6 Conclusions

In this Chapter, we proposed a new algorithm called Dual C -Means, which can be seen as an extension of the classical K -Means for dual representation spaces. Its originality lies in the fact that the clustering process can be driven by external resources by defining two distinct representation spaces. In particular, we proposed that query logs are used as external information to guide clustering and offer meaningful labels to users in their search for information²⁷. Based on previous findings that suggest the capabilities of higher p values, Dual C -Means only restrict the cluster centroid to p words and use as many words as possible to represent the Web snippets. Another advantage of this generalization is that, when query logs are used, p becomes a query log property instead of an initial parameter reducing the complexity of our proposal.

We also built a new publicly available dataset called WEBSRC401 based on ClueWeb09, which affords a more realistic situation for Web SRC. A complete and reproducible evaluation was performed over different gold standard datasets (ODP-239 and SEMEVAL) based on different publicly available evaluation tools. In particular, a great deal of evaluation metrics have been applied over different configurations of the Dual C -Means integrating distinct word-word similarity measures. Results have shown that our approach steadily outperforms all existing state-of-the-art SRC algorithms in terms of clustering accuracy (F_{b^3}). This result is due to the introduction of query logs, which allows high labeling quality with outperforming values of $I - rec@10$, $D - nDCG@10$ and $D\# - nDCG@10$.

The final findings that show that collocation metrics sensitive to high frequency events tend to produce high quality clusters and low frequency sensitive ones give rise to quality labels, is an issue with potential impact. Indeed, like the dual representation space, it suggests a multi-objective implementation of the dynamic reallocation algorithm to the problem of SRC²⁸.

Publications

This chapter has been validated by the following publications [Moreno et al., 2014, Moreno and Dias, 2013a]:

Moreno, J. G., Dias, G., and Cleuziou, G. (2014). Query log driven web search results clustering. In Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR), pages 777–786.

Moreno, J. G. and Dias, G. (2013). HULTECH at the NTCIR-10 INTENT-2 Task: Discovering User Intents through Search Results Clustering. In Proceedings of the 10th NII Testbeds and Community for Information Access Research Workshop (NTCIR), pages 140–146.

²⁷As shown in Chapter 2.

²⁸We discuss more about this idea in the Section 6.2.

Chapter 5

Web Image Clustered Visualization in a Mobile Context

Contents

4.1	Introduction	59
4.2	Dual C-means Algorithm	61
4.2.1	General Model	61
4.2.2	Instantiation in the SRC Context	62
4.3	The WEBSRC401 Dataset	65
4.4	Clustering Evaluation	66
4.4.1	Evaluation of SRC	66
4.4.2	Experimental Setups	68
4.4.3	Clustering Results	70
4.5	Labeling Evaluation	73
4.6	Conclusions	76

5.1 Introduction

In recent years, the growing number of mobile devices with Internet access has changed the way to access Web contents as well as user interaction [Kamvar and Baluja, 2006]. However, performing mobile Web image search is still made in a similar way as in desktop computers, i.e. a simple list or grid of ranked image results is returned to the user. Previous works on human-computer interaction have also shown that mobile user needs are different than the ones for desktop computers [Kamvar et al., 2009]. In particular, ranked lists are not suitable for exploration and selection of relevant results on mobile devices as they involve repeated scrolling, sliding and zooming.

In the specific context of Information Retrieval, [André et al., 2009] proposed a study of novel interfaces for Web image search. In particular, they conducted a large scale analysis of search logs based on a set of 55 million queries. Their findings suggest several interesting implications. Web image searchers view more pages of search results, they spend more time looking at those pages and they click on more results than Web page searchers. According to the authors, one of the main reasons for this situation is the fact that there is often no definitive answer to a query, which means that the sought after image could be one of many.

To remedy this situation, one common way to present Web search results on mobile devices is to build meaningful ephemeral clusters [Carpineto and Romano, 2009]. Outside the mobile computing context, this methodology has mainly been used for Web page search results organisation [Ferragina and Gulli, 2008] [Carpineto and Romano, 2009] [Carpineto and Romano, 2010] [Scaiella et al., 2012] and Web image search results exploration [Cai et al., 2004] [Wang et al., 2004] [Ding et al., 2008].

In this chapter, we are particularly interested in studying Web image SRC algorithms to improve search engine interfaces for mobile devices. Although different approaches have been proposed for Web image SRC [Cai et al., 2004, Wang et al., 2004, Ding et al., 2008], few studies have specifically been dealing with mobile devices [Liu et al., 2004, Moreno and Dias, 2011, Tolchinsky et al., 2012]. Moreover, both studies draw approximative conclusions as their evaluation frameworks are incomplete. As a consequence, we propose to evaluate the trade-off between clustering accuracy and used space-interface over a public dataset of 71478 Web images [Krapac et al., 2010] and 353 text queries for common baselines state-of-the-art SRC algorithms and for the algorithms introduced in previous chapters: Suffix Tree Clustering (STC) [Zamir and Etzioni, 1998], LINGO [Osinski et al., 2004], STC-LINGO, *AGK*-means and Dual *C*-means.

In the context of mobile devices, we hypothesize that SRC systems should successfully combine two main criteria: maximum cluster accuracy and minimum wasted space-interface¹. The underlying idea is simple. In order to leverage the users' efforts, a given interface should clearly list all of the many sought of images as well as present them in a compact representation. For the first case, ephemeral clustering should be as accurate as possible. For this purpose, we propose a broad set of clustering evaluation metrics [Amigó et al., 2009] in the specific context of text-based Web image SRC systems. For the second case, the diversity of the Web image search results should be presented in an effortless interface, which limits repetitive scrolling, sliding or zooming. In order to quantitatively measure the compactness of a given interface, we propose a new metric, which evaluates the mismatch of the used space-interface between the ground truth and the cluster distribution obtained by ephemeral clustering. The results evidence that there exist high divergences between clustering accuracy and used space maximization. As a

¹Label accuracy is also important, but it is studied in Section 4.5.

consequence, the trade-off of cluster-based exploration of Web image search results on mobile devices is difficult to define, although our study evidences some clear positive results.

5.2 Text-Based Web Image Search Results Clustering

Web image SRC systems consist in different processing steps to address the organization of Web image search results following two different approaches: single-step and multiple-steps. The common processing in both approaches is text-based Web image search results clustering. Indeed, both methodologies strongly depend on (text) Web image snippet ephemeral clustering, which may eventually be combined with other features. As a consequence, we propose to study different text-based SRC algorithms. Related studies in Web image SRC have privileged two different clustering algorithms: [Zamir and Etzioni, 1998, Osinski et al., 2004]. However, many successful works have been proposed for ephemeral text clustering such as *AGK*-means and Dual *C*-means². As a consequence, we propose a comparative study to acknowledge the behavior of each one of these algorithms in terms of a new evaluation metric to measure automatically the users' effort.

5.2.1 A Gallery-Based Interface

Within the scope of our study, we also propose a gallery-based interface to present the results obtained by Web image SRC algorithms. Under this interface, each cluster is displayed with its label and a gallery of Web images that belongs to it. So, each cluster of image result can be explored with left-right movements as in a typical gallery exploration (see Figure 5.1). Moreover, shift between clusters is performed using up-down movements allowing a quick exploration of different meanings/facets discovered by the SRC algorithm. Note that this interface takes advantage of touch-screen capabilities allowing the smooth integration of clusters in mobile interfaces³. With this interface, the users can explore the top image results in the first clusters like in a common grid interface. However, as additional feature, they can explore the next image results of potentially interesting clusters depending on their query intentions. This interface offers two main advantages for cluster-based interfaces: user adaptability (because of the use of widely-known interactions through the use of a gallery interface) and navigation (because when the shift between clusters is performed, the state of the gallery is kept and the users can restart the exploration of previously explored clusters in the deserted point). Although the interface implementation is thought to minimize users' efforts in both scrolling and

²These two algorithms underperform other studies such as OPTIMSRC [Carpineto and Romano, 2010] and TOPICAL [Scaiella et al., 2012].

³Note that the same interface can be used through multidirectional button or keyboard-based navigation.

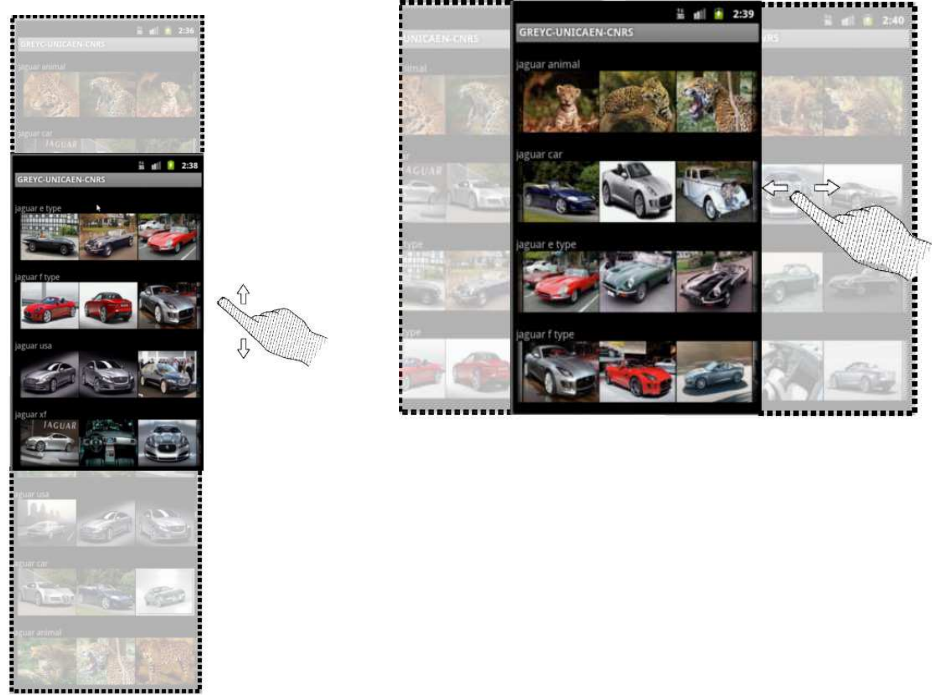


Figure 5.1: Mobile Interface. (Left) Mobile Interface, (Center) Scroll Interface Operation and (Right) Slide Interface Operation.

sliding, it is important to note that what determines the “real” user energy to use this kind of interfaces is the quantification of the average number of images in each cluster and the average number of clusters to display. We will discuss this issue in the next section.

5.3 The Web Image SRC Dataset

To evaluate SRC algorithms in the context of mobile Web image search, we use a public dataset of 71478 images proposed by [Krapac et al., 2010] composed by the top Web image results for 353 text queries. For each Web image result, its associated metadata are available. Each (text) Web image snippet is constructed from the text surrounding the image (10 words before and after), the title of the Web page and the alternative image description, available through the metadata. For each Web image, a binary label is included to assess if it is relevant or irrelevant to the query. However, this dataset is not directly usable as a SRC dataset. For this reason, we use the golden standard to built an intermediate dataset composed by only two classes of images for each query (relevant or irrelevant).

However, for the sake of our evaluation, we need to extend the queries to ambiguous/multi-faceted cases. As such, we propose to merge related queries in terms of string matching and generate a list of new queries. The construction of this new dataset is described by the following steps: (1) each query is tokenized and the frequency of each unique token is calculated, (2) unique tokens with frequency higher than two are selected as new queries, (3) the image list of each new query is determined by the union of all the image lists of the queries in which the new query is present and (4) the new query membership is defined in a similar way as in the original dataset. Following this procedure, a total of 61 new ambiguous/multi-faceted queries are obtained and the number of clusters for each new query varies between 4 (when 2 queries are merged) and 50 (when 25 queries are merged). Note that the obtained dataset has been formatted to the standard proposed by [Carpineto and Romano, 2010]. We will refer to it as the Web image SRC dataset⁴. Some examples of the new dataset are illustrated in Table 5.1.




New Query	Merged Queries (Amb./Facets)	Web Image Results Examples
logo	logo psg, logo fc barcelona, logo apple, logo windows, logo renault, logo ferrari, logo adidas, logo nike, etc.	
france	stade de france, france flag, map france, france team jersey	
simpson	bart simpson, homer simpson	

Table 5.1: Examples of the Web image SRC dataset for ambiguous/multi-faceted queries.

5.4 Evaluation and Results

5.4.1 Clustering Performance

Following previous evaluation setups, we evaluate the clustering performance of the state-of-the-art algorithms against our solutions, e.g. the *AGK*-means⁵ and Dual *C*-means algorithms. The clustering results - using F_1^C , F_1^N and F_{b3} - are presented in Tables 5.2 and 5.3. Obtained results for the Web image dataset are consistent with the results obtained in previous chapters, e.g. the Dual *C*-means configurations outperform many

⁴This dataset is available upon request to the authors.

⁵Using only SCP.

of the other algorithms with some exceptions⁶. For this reason, our experiments will be more concentrate in the analysis of the wasted space interface metric.

Dataset	Metric	STC	LINGO	BiKm	AGK-means
Web image SRC	F_1^C	0.469	0.247	0.414	0.553
	F_1^N	0.534	0.390	0.472	0.574
	F_{b^3}	0.475	0.267	0.414	0.554

Table 5.2: Ephemeral clustering results using the Web image SRC dataset.

Dataset	Metric	Dual C-means				
		SCP	PMI	DICE	LOGLIKE	Φ^2
Web image SRC	F_1^C	0.596	0.486	0.598	0.563	0.555
	F_1^N	0.587	0.527	0.591	0.571	0.567
	F_{b^3}	0.606	0.499	0.607	0.572	0.563

Table 5.3: Ephemeral clustering results for Dual C-means using the Web image SRC dataset.

5.4.2 Wasted Space-Interface

In this work, we strongly believe that the diversity of the Web image search results should be presented in an effortless interface, which can limit repetitive scrolling, sliding and zooming. But, objectively measuring this issue in mobile devices is still an open problem. In order to address this issue, a new metric called Wasted Space-Interface (WSI) is proposed. The idea behind is simple: under equal display sizes, the less (interface) space is used to present a given quantity of information, the more the users' efforts will be leveraged. So, the WSI should evaluate the mismatch of the used space-interface between the ground truth and the cluster distribution obtained by ephemeral clustering. One direct implication is that the measure does not depend on the physical characteristics of the mobile device, e.g. screen size, but is correlated to the query results distribution in the gold standard.

For that purpose, we first define two different spatial quantities: A_{gsr} and A_{src} in Equations 5.1 and 5.2, respectively.

$$A_{gsr} = GS_{max} \times n. \quad (5.1)$$

The area used by each gold standard query result (A_{gsr}) is defined by the product between the number of elements of the cluster with more images ($GS_{max} = \max(|L_1|, |L_2|, \dots, |L_n|)$, where $|L_i|$ is the number of images in each cluster L_i) and the number of clusters (n). Note that this area is related to the number of repetitive scrolls

⁶Remember that results obtained with AGK-means are closely to Dual C-means when SCP is used as collocation measure.

and slides to explore the overall results. Indeed, the exploration of the images in a given cluster is related to sliding and the shift between clusters is related to scrolling (or vice and versa depending of the screen position).

$$A_{src} = CS_{max} \times m. \quad (5.2)$$

Correspondingly, the area used by any SRC algorithm for each query result (A_{src}) is the product between the size of the biggest discovered cluster ($CS_{max} = \max(|C_1|, |C_2|, \dots, |C_m|)$) and the number of produced clusters (m). Finally, the WSI is defined in Equation 5.3, and it corresponds to the difference between the used area-interface of any SRC algorithm minus the used area-interface by the gold standard.

$$WSI = \max\{A_{src} - A_{gsr}, 0\}. \quad (5.3)$$

Note that, if the golden standard clustered results were presented for a given query on an imaginary screen, the WSI metric would evaluate the need for an extra screen size to display all the information present in the results⁷. Therefore, less compact SRC algorithms would include additional users' effort as repetitive interface interactions would be necessary to visualize all information. As a consequence, more compact SRC algorithms are more likely to afford less repetitive scrolling, sliding and zooming and accordingly better user explorations of the clustered Web image results can be achieved. To illustrate this situation, we present in Figures 5.2 and 5.3 the discovered clusters distributions for an example query over the standard dataset for each one of the five SRC tested algorithms. In particular, each circle represents a Web image result, the cluster membership is determined by the horizontal organization (each line is a different cluster) and the membership to the gold standard clusters is represented by different colors.

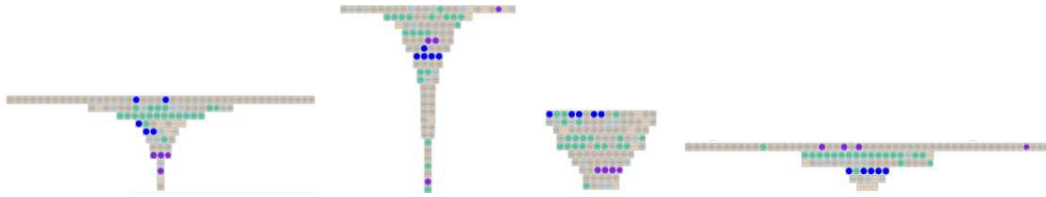


Figure 5.2: Distribution shapes for SRC algorithms. From left to right: STC, LINGO, BiKm and AGK-means.

To allow direct comparison of the WSI metrics between different queries, we define the Normalized WSI (NWSI). In particular, we take into account the fact that the worst SRC algorithm would maximize the A_{src} value and therefore NWSI is obtained by dividing WSI with A_{src}^{max} . It is easy to prove that $A_{src}^{max} = ((n_{q_j} + 1)/2)^2$, where n_{q_j} is the number of images of the gold standard for any query (q_j). As a consequence, the NWSI metric is

⁷The negative extra space is avoided by the zero value in the definition of WSI.

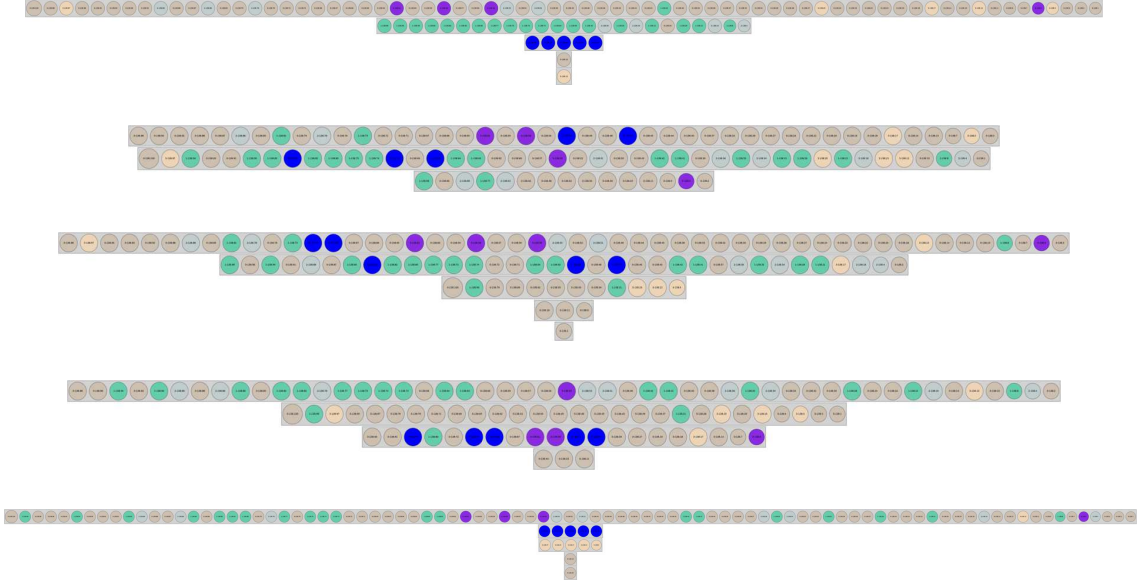


Figure 5.3: Distribution shapes for Dual C -means. From top to bottom: SCP, PMI, LOGLIKE, Φ^2 and DICE.

defined as $NWSI = WSI/A_{src}^{max}$. The $NWSI$ is completely defined in Equation 5.4 and the normalized results of the wasted space-interface are shown in Table 5.4.

$$NWSI = \frac{WSI}{((n_{q_j} + 1)/2)^2}. \quad (5.4)$$

Algorithm		NWSI
STC		0.082
LINGO		0.305
BiKm		0.031
AGK-means		0.027
Dual C -means	SCP	0.026
	PMI	0.019
	DICE	0.027
	LogLike	0.017
	Φ^2	0.030

Table 5.4: Average $NWSI$ values for the overall queries in the image SRC dataset.

Note that the smaller the $NWSI$ is, the less space is needed to present the clustered information and therefore users' efforts are minimized. The results clearly demonstrate that the $NWSI$ is smaller on average for the AGK-means than the classical SRC algorithms. Moreover, all of the configurations of the Dual C -means are capable to improve the $NWSI$ quality obtained against the AGK-means. This result is not surprising given that Dual C -means is a generalisation of AGK-means and that the dataset does not permit the use

of dual spaces⁸. Moreover, both algorithms provide a more compact representation of the clustered information and confirm our second hypothesis.

Overall results evidence that there exist high divergences between clustering accuracy and used space-interface maximization for classical SRC algorithms. For cluster accuracy, we base our conclusions on the F_{b3} results following the recommendations of [Amigó et al., 2009]. Within this context, the *AGK*-means algorithm outperforms by 17.7% the second best result achieved by the *STC* algorithm. The same situation can be observed for the *NWSI* evaluation as best results are obtained for the *AGK*-means, although comparative results are obtained for *BiKm*. In fact, these results indicate that even though *BiKm* does not achieve as good performances as *STC* in terms of cluster accuracy, it is more suitable for interface definition. Comparatively, the results for *LINGO* show bad performances for both experiments. As such, our study clearly and exhaustively demonstrates that the state-of-the-art SRC algorithm, *AGK*-means, achieves good performances both in terms of clustering accuracy and used space-interface. These results indicate that current SRC algorithms are appropriate to deal with Web image SRC exploration for mobile devices.

5.5 Conclusions

In this Chapter, we present the first exhaustive evaluation for the design of enhanced Web image SRC systems on mobile devices. We base our study on two different hypotheses. First, cluster accuracy must be maximized and second used space-interface is ought to be maximized. For that purpose, we develop a complete evaluation framework based on (1) the definition of a new dataset for multi-faceted Web image search from the original dataset proposed in [Krapac et al., 2010], (2) the use of newly introduced clustering evaluation metrics presented in [Amigó et al., 2009] and (3) the definition of the Normalized Wasted Space-Interface, which evaluates the users' effort to explore Web search results based on a spatial definition that does not rely on hardware specificities. The results evidence that there exist divergences between cluster accuracy and *NWSI* for "classical" SRC algorithms. However, state-of-the-art algorithms for text-based ephemeral clustering, such as Dual *C*-means and *AGK*-means, propose the best trade-off between accuracy and usage. As a consequence, new SRC algorithms are well-suited to generate compact interfaces for mobile devices, thus inherently reducing scrolling, sliding and zooming.

⁸It is important to remark that Dual *C*-means can be used with or without external resources. In order to use this image dataset, the set of experiments were conducted using the only text configuration, i.e. Dual *C*-means with $V_1 = V_2$.

Publications

This chapter has been validated by the following publications [Moreno and Dias, 2013b, Moreno and Dias, 2012]:

Moreno, J. G. and Dias, G. (2013). Using text-based web image search results clustering to minimize mobile devices wasted space-interface. In Proceedings of the 35th European conference on Advances in Information Retrieval (ECIR), pages 532–544.

Moreno, J. G. and Dias, G. (2012). Image Results Exploration using Ephemeral Clustering. Demo Paper Selected by the CNRS for research center representation at World Wide Web Conference (CNRS-WWW), pages 1-3.

Chapter 6

Conclusion and Future Work

Contents

5.1	Introduction	77
5.2	Text-Based Web Image Search Results Clustering	79
5.2.1	A Gallery-Based Interface	79
5.3	The Web Image SRC Dataset	80
5.4	Evaluation and Results	81
5.4.1	Clustering Performance	81
5.4.2	Wasted Space-Interface	82
5.5	Conclusions	85

6.1 Conclusions and Contributions

As mentioned in the introduction of this thesis, during this work, the problem of visualizing Web image search results has been addressed using different approaches including several algorithms. Main findings were developed and presented to two separated but related communities, IR and NLP. First, the search results clustering problem is addressed from the text point of view starting with the baselines and obtaining as a result novel search results clustering strategies. This is a hot topic in IR and it is reflected by the high rate of recent published papers. On the other hand, NLP conferences are also interested in this subject because the application of SRC algorithms is a challenging task which may include word sense disambiguation and entity resolution.

On the initial chapters, we presented the definition of the problem and the corresponding state-of-the-art algorithms. Towards the middle of the thesis, we present our proposed algorithms and they are tested with known and particular developed datasets in order to ensure a complete evaluation including unexplored aspects in SRC as: clustering quality, labeling quality and users' efforts. Our results strongly support our hypothesis about

the capabilities of the novel methods showing that an important improvement can be achieved in comparison with classical and strong state-of-the-art algorithms. Therefore, it is demonstrated that our solutions are suitable in the explored problem and under the mobile device restrictions. Our new proposed datasets, the WEBSRC401 and WEBImageSRC dataset, include a set of peculiarities that creates new challenges that could be tackled by future research. The presented evaluation includes rigorous protocols with fully defined and reproducible metrics. However, the reproducibility is a challenge within any research project that includes some limitations as it is discussed in Section 1.5.

In this thesis, several contributions have been obtained as a result of our dedicated effort. Indeed, the contributions of this thesis can be separated in three groups: methodologies, research datasets and research tools.

As for methodologies, we have developed the *AGK*-means and the Dual *C*-means algorithms. Each of them correspond to a different behaviour, but the latter one can be seen as a generalisation of the early one. The most interesting feature of Dual *C*-means, it is the possibility to integrate external resources in the labeling process which allows to preserve the good performance obtained by *AGK*-means in terms of clustering but it allows the use of more natural labels for the clusters. Capabilities of this algorithm are evaluated in many aspects including clustering quality, labeling quality and users' effort. Obtained results support our statement that Dual *C*-means is a suitable algorithm to use in Web images exploration on mobile devices.

As for datasets, we have built two of them based on two exists ones. First, the WEBSRC401 dataset described in Section 4.3 and second, the Web image SRC dataset described in Section 5.3. Both of them, are publicly available and follow the standard format widely accepted in the SRC community. Each dataset has different evaluation purposes. WEBSRC401 allows clustering evaluation as well as labeling evaluation when combined with the Intent2 NTCIR dataset¹. On the other hand, the Web image SRC dataset allows clustering and users' effort evaluation which permits the measuring of this interesting user behavior in an automatic way. Limitations regarding these datasets are similar to the known alternatives. However, our proposed datasets are the most complete options when automatic and reproducible evaluation is desired.

Finally, as for research tools, we have proposed a new evaluation tool for the SRC problem. In it, we have implemented a considerable number of traditional metrics as well as recent metrics proposed by other authors or in our papers. This tool follows a recently proposed standard output in order to facilitate the posterior comparison phase and the use of alternative evaluation tools. In order to motivate new research studies, the strong baselines were developed in an easy-to-use tool with the possibility of free configuration. Following the open research spirit, both tools are publicly available in the form of jar files. Our main developed project, including the Dual *C*-means algorithm, is available under

¹Further information about this dataset could be find in Section 4.3

request.

We have presented all these results as research papers in different conferences. At the end of each chapter were mentioned the paper used to validate the presented results. However, the full list of the publications - enumerated by chapter - achieved during the developing of this thesis is mentioned below:

- Chapter 2

Moreno, J. G. and Dias, G. (2011). Using ephemeral clustering and query logs to organize web image search results on mobile devices. In Proceedings of International ACM workshop on Interactive Multimedia on Mobile and Portable Devices (IMMPD), pages 33–38.

Moreno, J. G. and Dias, G. (2014). Easy web search results clustering: When baselines can reach state-of-the-art algorithms. In Proceedings of 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pages 1–5.

- Chapter 3

Moreno, J. G., Dias, G., and Cleuziou, G. (2013). Post-retrieval clustering using third-order similarity measures. In Proceedings of 51st Annual Meeting of the Association for Computational Linguistics (ACL), pages 153–158.

- Chapter 4

Moreno, J. G., Dias, G., and Cleuziou, G. (2014). Query log driven web search results clustering. In Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR), pages 777–786.

Moreno, J. G. and Dias, G. (2013). HULTECH at the NTCIR-10 INTENT-2 Task: Discovering User Intents through Search Results Clustering. In Proceedings of the 10th NII Testbeds and Community for Information Access Research Workshop (NTCIR), pages 140–146.

- Chapter 5

Moreno, J. G. and Dias, G. (2013). Using text-based web image search results clustering to minimize mobile devices wasted space-interface. In Proceedings of the 35th European conference on Advances in Information Retrieval (ECIR), pages 532–544.

Moreno, J. G. and Dias, G. (2012). Image Results Exploration using Ephemeral Clustering. Demo Paper Selected by the CNRS for research center representation at World Wide Web Conference (CNRS-WWW), pages 1–3.

- Chapter 6

Moreno, J. G. and Dias, G. (2014). PageRank-based Word Sense Induction within Web Search Results Clustering. In Proceedings of the 14th ACM/IEEE Joint Conference on Digital Libraries (JCDL), pages 1-2.

Acharya, S., Saha, S., **Moreno, J. G.** and Dias, G. (2014). Multi-Objective Search Results Clustering. In Proceedings of the 25th International Conference on Computational Linguistics (COLING). pages 99–108.

To summarize, in this thesis, we present a solution for the Web image search results clustering particular to mobile devices. The presented solution uses common interfaces in a new way in order to allow new possibilities when a user interacts with the Web results. Our proposed interface is also inspired by the current characteristics of mobile devices as in smartphones or tablets. The proposed interface allows parallel exploration of Web image results and our experiments support our idea that users' effort is reduced when appropriate SRC algorithms are used.

After exhaustive evaluation, following are our main findings:

- A complete evaluation of commonly used baselines with recent clustering evaluation metrics.
- Novel combination of traditional baselines to understand their maximum capabilities.
- First SRC algorithm that relies on collocation measures as the similarity metric with automatic cluster size definition.
- Our novel extension of the K -means algorithm, the Dual C -means algorithm, outperforms the actual existing state-of-the-art algorithms in terms of cluster quality and also achieves competitive label quality when compared with strong baselines.
- Label driving clustering algorithms preserve label quality without loss in cluster quality.
- New dataset and evaluation framework that automatically evaluates of SRC labels.
- In terms of user effort, our proposal outperforms existing strategies, which support our hypothesis that compact representation of the results reduce the exploration effort.

Our proposed research questions are addressed through out of this thesis. The initial hypothesis that association metrics could be useful in the improvement of clustering quality for ephemeral clustering algorithms is verified with our experiments, which were particularly evaluated in the Web image exploration task. A new mobile interface is exploited transversely thorough this thesis with interesting results in terms of the three main addressed challenges: clustering accuracy, labeling quality and user effort. The results of

this thesis allow new research directions and two of our explored preliminary possibilities are presented in the following section.

6.2 Preliminary studies in Future Directions

In the short future, two main ideas could be addressed. Indeed, they are not mutually exclusive and an interesting direction is their combination with more complex algorithms. The most straightforward is the integration of multi-objective algorithms that consider all the explored aspects as clustering quality, labeling quality and users' effort. In order to explore this possibility, we have included our preliminary results following this direction. However, we consider that a complete solution clearly demands a big research effort. On the other hand, one more out-of-the-box solution is also explored. In this case, content information is completely ignored and the similarity is defined only by interlink analysis. Obtained results show that this is a promising alternative, which rates into account only Web structure mining techniques.

6.2.1 A Novel Non-Content Based Direction

In general, SRC has been addressed as a content based clustering problem, i.e., Web snippet content is exploited to identify relevancy between documents. In this Section, we explore an alternative direction. It is based on the idea that structural information of the Web could reveal information about the content. Similar ideas have been explored previously in other studies [Avrachenkov et al., 2008], but none has addressed the SRC task.

The underlying idea is simple and grounded on the following hypothesis: word senses are distributed over the Web in the same way Web pages are linked together. In other words, Web pages containing the same information related to one topic should share some similar linking properties. This hypothesis supposes that (1) meanings are separated by linking importance of the Web and (2) Web domains provide a unique meaning of a given word, thus extrapolating the "one sense per discourse" paradigm defined by [Gale et al., 1992]. So, if both factors are true, performing clustering over link-based similarity metrics could be a valid SRC solution.

PageRank-based Search Results Clustering PageRank-based clustering has proved to be a useful strategy for hypertext document clustering [Avrachenkov et al., 2008]. So, we adapt these ideas to SRC as clustering is run over the sub-collection returned by the search engine and not over the entire Web collection. As a consequence, we propose to use the Jensen-Shannon Divergence metric to calculate similarities between hypertext

documents². To calculate the kernel values between two documents, we use the Jensen-Shannon kernel proposed by [Martins et al., 2009]. As a clustering algorithm, we have chosen the Spectral Clustering Algorithm³ where each output cluster is considered as one unique sense and evaluated as it.

Results and Discussion We have used the Hyperlink Graph dataset publicly available and defined in [Meusel et al., 2014] to calculate the PageRank values of the SEMEVAL dataset. Even when PageRank clustering algorithms does not get the top position, the results indicate that linking information is useful in the word sense disambiguation task. The obtained result ($F_1^N = 0.6367$) is not far from the value obtained by LDA ($F_1^N = 0.7159$).

Conclusive Remarks We have presented a SRC algorithm based on PageRank clustering. Results show that non-content-based clustering algorithms can achieve competitive results when compared with content-based. In consequence, the PageRank clustering algorithm allows us to capture each sense in each cluster. An interesting direction could be the combination between content and non-content-based strategies which could allow the improvement of the overall performance of the SRC systems. For example, we could follow a multi-objective solution.

6.2.2 A Multi-objective Search Results Clustering Direction

Recent advances show that it is possible to perform clustering through the use of multiple objectives or constraints [Morik et al., 2012, Métivier et al., 2012]. Indeed, a successful algorithm used in the Web image search results clustering problem must consider different factors. If an objective function can be defined for each of these factors, they could be addressed with a multi-objective (MOO) strategy. In this is exploratory direction, our intention is not implementing all objectives to completely cover the entire set of restrictions, but instead we propose a simple alternative clustering solution of the Dual *C*-means algorithm with a MOO clustering strategy.

Clustering as a MOO Problem As far as we know, within text applications, [Morik et al., 2012] is the first work, which formulates text clustering a multi-objective optimization problem. In this work, we follow their main ideas combined with the AMOSA framework [Bandyopadhyay et al., 2008]. Multi-objective optimization can be formally stated as finding the vector of decision variables that simultaneously optimize a

²[Avrachenkov et al., 2008] discarded this option as it was computationally expensive in their reserach work over the entire Web.

³Implemented in SciKit Learn tool <http://scikit-learn.org/> [Last access: 11/06/2014].

set of objective function values while satisfying user-defined constraints, if any. In general, a MOO algorithm outputs a set of solutions not dominated by any solution encountered by it. Within the specific context of clustering, two objective functions are usually defined, which must be optimized simultaneously: *Compactness* of the documents in a cluster and *Separability* between the centroids of the clusters. These two functions were defined using derivations of the Equation 4.1.

Results and Discussion We evaluated our algorithm with ODP-239 dataset using the SCP collocation measure and different p values for the centroid definition in Equation 3.2. The proposed solution evidences a marginal sensitivity to different p values. Indeed, for ODP-239, changing p between 2 and 5 words has a negligible impact on F_{b3} . So, p can be seen as a non influential parameter for clustering purposes. Using the best configurations, the MOO solution achieves a score of $F_{b3} = 0.484$ which outperforms the *AGK*-means result and slightly approach the result obtained by Dual *C*-means of $F_{b3} = 0.485$.

Conclusive Remarks It is the first attempt to define the SRC task as a MOO problem. Even when multiple objectives are explored, this solution not consider all the collocation measures studied in this theses. However, its open a new branch of possibilities in SRC. A complete combination with the all the characteristics of the Dual *C*-means algorithm could improves the obtained solutions in terms of cluster and labeling quality. An important additional characteristics of a MOO solution, is that it could consider non-content information as proposed in Section 6.2.1.

6.3 Final Remarks

In this chapter, we summarized our studies on ephemeral clustering techniques for Web image results exploration and additionally, we proposed future research directions with preliminary results. Our findings suggest some future intensive studies around the combination of different clues and factors which are possible do with the Dual *C*-means algorithm and the multi-objective strategy, either through the combination of content based and non-content based strategies or the combination of different collocation measures. It is important to remark that, when are used within Dual *C*-means, some collocation measures (such as SCP) tend to provide good clusters, but some others (such as PMI) tend to improve labeling quality. In this case, the use of MOO seems to be an attractive idea. However, computational efficiency is an important issue present in the actual solutions for MOO clustering.

On the other hand, our proposed solution, Dual *C*-means, is an suitable solution for the Web image exploration problem. However, combination with other external resources

could be investigated, but additional efforts in terms of evaluation must be need. Another challenging experiment to perform, it is the evaluation by users of our developed algorithms when Web image search are performed. Many alternative directions are likely to explored because of the accomplishments of this thesis.

Publications

Sections 6.2.1 and 6.2.2 have been validated by the following publications [Moreno and Dias, 2014b, Acharya et al., 2014]:

Moreno, J. G. and Dias, G. (2014). PageRank-based Word Sense Induction within Web Search Results Clustering. In Proceedings of the 14th ACM/IEEE Joint Conference on Digital Libraries (JCDL), pages 1-2.

Acharya, S., Saha, S., **Moreno, J. G.** and Dias, G. (2014). Multi-Objective Search Results Clustering. In Proceedings of the 25th International Conference on Computational Linguistics (COLING). pages 99–108.

List of Figures

1.1	Search of the query “android” in amazon.com [Last access: 17/06/2014].	8
1.2	Search of the query “android” in carrot2.org [Last access: 17/06/2014].	9
2.1	Web image results for the query “colombia” in the Android platform.	19
2.2	Results for the query “jaguar” using Ephemeral Clustering (HISGK-means [Dias et al., 2011]) and Query Logs.	24
2.3	Percentage of users agreement for Text, Images and Combined questions from 45 AMTurk workers.	28
2.4	Estimated Average Accumulative results over the 97 queries for the first 10 image groups.	29
2.5	Boxplot for percentage of overlapping between the first 64 results of the original query and the first 64 results of the expanded queries.	30
2.6	Representative works of SRC over time: STC [Zamir and Etzioni, 1998], LINGO [Osinski and Weiss, 2005], OPTIMSRC [Carpineto and Romano, 2010], TOPICAL [Scaiella et al., 2012] and G-NGRAMS [Di Marco and Navigli, 2013].	32
2.7	F_{β}^C and F_{b3} for Moresque and ODP-239 for Random Clustering.	41
2.8	F_1^C and F_{b3} for Moresque and ODP-239 for Standard and Tuned Clustering.	41
3.1	$J_{S_{3rd}}$ and its modelisation.	51
3.2	Values of β (on the left) and differences between consecutive values of β (on the right).	53
4.1	Dual C -Means aims to discover clusters of objects in E_1 closed to a common cluster representative in E_2 .	61
4.2	Example of the Dual C -Means instantiated for the SRC context with external information labels as cluster label space.	64
4.3	Impact of K for F_{b3} against ODP-239 dataset.	71
4.4	Impact of K for F_{b3} against WEBSRC401 dataset.	71
5.1	Mobile Interface. (Left) Mobile Interface, (Center) Scroll Interface Operation and (Right) Slide Interface Operation.	80

- 5.2 Distribution shapes for SRC algorithms. From left to right: STC, LINGO, BiKm and *AGK*-means. 83
- 5.3 Distribution shapes for Dual *C*-means. From top to bottom: SCP, PMI, LOGLIKE, Φ^2 and DICE. 84

List of Tables

2.1	Categorization, number and queries used for the evaluations.	27
2.2	Clustering Evaluation Metrics.	35
2.3	Description of the SRC gold standard datasets.	37
2.4	State-of-the-art results using F_1^C for several SRC algorithms. (*) See table footnote.	37
2.5	Standard, Tuned and Random Results for Moresque dataset.	38
2.6	Standard, Tuned and Random Results for ODP-239 dataset.	38
2.7	Cascade Results for Moresque datasets.	39
2.8	Cascade Results for ODP-239 datasets.	40
3.1	F_{b3} for <i>SCP</i> for the global search and the stopping criterion for the ODP-239 dataset.	55
3.2	F_{b3} for <i>PMI</i> for the global search and the stopping criterion for the ODP-239 dataset.	55
3.3	SRC comparative results for F_β^C and F_{b3} over the ODP-239 dataset.	56
3.4	SRC comparative results for F_β^C and F_{b3} over the Moresque dataset.	57
4.1	Text query and associated subtopics for the queries ids 155 and 160 in the WEBSRC401 dataset. The Web snippets were selected from the list of relevant documents that were manually annotated in the TREC Web Track 2012.	67
4.2	Results of the state-of-the-art algorithms for ODP-239, SEMEVAL and WEB-SRC401. K was fixed to 10 clusters for TOPICAL and LDA. Other algorithms were ran using standard configurations.	72
4.3	Results of the Dual <i>C</i> -Means algorithm for ODP-239 and SEMEVAL. K fixed to 10 Clusters. Let us notice that for all experiments, the number of p words composing the centroids was set to 2 and the vocabulary is the set of words appearing in the retrieved Web snippets.	72

4.4	Results of the Dual C -Means algorithm for WEBSRC401. K was fixed to 10 clusters. Note that for the experiments when $V_1 = V_2$, the number of p words composing the centroids was set to 2 and the vocabulary is the set of words appearing in the retrieved Web snippets. Note that \dagger means paired student's t-test statistical relevance for $p - value < 0.05$ between a given metric in $V_1 = V_2$ and its counterpart in $V_1 \neq V_2$	73
4.5	Evaluation results of the labeling process with query logs over the NTCIR-10 Intent-2 dataset. Note that \dagger means paired student's t-test statistical relevance for $p - value < 0.05$ between a given metric in $V_1 = V_2$ and its counterpart in $V_1 \neq V_2$	75
5.1	Examples of the Web image SRC dataset for ambiguous/multi-faceted queries.	81
5.2	Ephemeral clustering results using the Web image SRC dataset.	82
5.3	Ephemeral clustering results for Dual C -means using the Web image SRC dataset.	82
5.4	Average NWSI values for the overall queries in the image SRC dataset. . .	84

Bibliography

- [Acharya et al., 2014] Acharya, S., Saha, S., Moreno, J. G., and Dias, G. (2014). Multi-objective search results clustering. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 99–108.
- [Aitken, 1926] Aitken, A. (1926). On bernoulli’s numerical solution of algebraic equations. *Research Society Edinburgh*, 46:289–305.
- [Amigó et al., 2009] Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486.
- [Amigó et al., 2013] Amigó, E., Gonzalo, J., and Verdejo, F. (2013). A general evaluation measure for document organization tasks. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 643–652.
- [André et al., 2009] André, P., Cutrell, E., Tan, D. S., and Smith, G. (2009). Designing novel image search interfaces by understanding unique characteristics and usage. In *Proceedings of 12th International Conference on Human-Computer Interaction (INTERACT)*, pages 340–353.
- [Arthur and Vassilvitskii, 2007] Arthur, D. and Vassilvitskii, S. (2007). K-means++: the advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1027–1035.
- [Avrachenkov et al., 2008] Avrachenkov, K., Dobrynin, V., Nemirovsky, D., Pham, S., and Smirnova, E. (2008). Pagerank based clustering of hypertext document collections. In *Proceedings of the 31st Annual International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 873–874.
- [Bahmani et al., 2012] Bahmani, B., Moseley, B., Vattani, A., Kumar, R., and Vassilvitskii, S. (2012). Scalable k-means++. *Proceedings of the Very Large Data Base Endowment (PVLDB)*, 5(7):622–633.

- [Bandyopadhyay et al., 2008] Bandyopadhyay, S., Saha, S., Maulik, U., and Deb, K. (2008). A simulated annealing-based multiobjective optimization algorithm: Amosa. *Transactions on Evolutionary Computation*, 12(3):269–283.
- [Bernardini et al., 2009] Bernardini, A., Carpineto, C., and D’Amico, M. (2009). Full-subtopic retrieval with keyphrase-based search results clustering. In *Proceedings of IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 206–213.
- [Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- [Brants and F., 2006] Brants, T. and F., A. (2006). Web 1t 5-gram.
- [Cai et al., 2004] Cai, D., He, X., Li, Z., Ma, W.-Y., and Wen, J.-R. (2004). Hierarchical clustering of www image search results using visual, textual and link information. In *Proceedings of 12th Annual ACM International Conference on Multimedia (MM)*, pages 952–959.
- [Carpineto et al., 2011] Carpineto, C., D’Amico, M., and Bernardini, A. (2011). Full discrimination of subtopics in search results with keyphrase-based clustering. *Web Intelligence and Agent Systems*, 9(4):337–349.
- [Carpineto et al., 2009] Carpineto, C., Osinski, S., Romano, G., and Weiss, D. (2009). A survey of web clustering engines. *ACM Computer Survey*, 41(3):1–38.
- [Carpineto and Romano, 2009] Carpineto, C. and Romano, G. (2009). Mobile information retrieval with search results clustering : Prototypes and evaluations. *Journal of the American Society for Information Science*, 60:877–895.
- [Carpineto and Romano, 2010] Carpineto, C. and Romano, G. (2010). Optimal meta search results clustering. In *Proceedings of 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 170–177.
- [Chatzichristofis et al., 2010] Chatzichristofis, S. A., Zagoris, K., Boutalis, Y. S., and Papatrakis, N. (2010). Accurate image retrieval based on compact composite descriptors and relevance feedback information. *International Journal of Pattern Recognition and Artificial Intelligence*, 24(2):207–244.
- [Church and Hanks, 1990] Church, K. and Hanks, P. (1990). Word association norms mutual information and lexicography. *Computational Linguistics*, 16(1):23–29.
- [Cutting et al., 1992] Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W. (1992). Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR)*, pages 318–329.

- [Datta et al., 2008] Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60.
- [Deerwester et al., 1990] Deerwester, S. C., Dumais, S. T., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.
- [Di Marco and Navigli, 2013] Di Marco, A. and Navigli, R. (2013). Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics*, 39(3):709–754.
- [Dias et al., 2007] Dias, G., Alves, E., and Lopes, J. (2007). Topic segmentation algorithms for text summarization and passage retrieval: An exhaustive evaluation. In *Proceedings of 22nd Conference on Artificial Intelligence (AAAI)*, pages 1334–1339.
- [Dias et al., 2011] Dias, G., Cleuziou, G., and Machado, D. (2011). Informative polythetic hierarchical ephemeral clustering. In *Proceedings of IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 104–111.
- [Dice, 1945] Dice, L. (1945). Measures of the amount of ecologic association between species. *Journal of Ecology*, 26:297–302.
- [Ding et al., 2008] Ding, H., Liu, J., and Lu, H. (2008). Hierarchical clustering-based navigation of image search results. In *Proceedings of 16th Annual ACM International Conference on Multimedia (MM)*, pages 741–744.
- [Downey et al., 2007] Downey, D., Broadhead, M., and Etzioni, O. (2007). Locating complex named entities in web text. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2733–2739.
- [Dunning, 1993] Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.
- [Ferragina and Gulli, 2008] Ferragina, P. and Gulli, A. (2008). A personalized search engine based on web-snippet hierarchical clustering. *Software: Practice and Experience*, 38(2):189–225.
- [Gale and Church, 1991] Gale, W. and Church, K. (1991). Concordances for parallel texts. In *Proceedings of the 7th Annual Conference of the UW Center for the New OED and Text Research, Using Corpora*, pages 40–62.
- [Gale et al., 1992] Gale, W., Church, K., and Yarowsky, D. (1992). One sense per discourse. In *Proceedings of the Workshop on Speech and Natural Language of the Human Language Technology Conference (HLT)*, pages 233–237.

- [Google, 2010] Google (2010). The New Image Search for Android and iPhone. [Online; accessed June-2014]. <http://googlemobile.blogspot.com/2010/04/new-image-search-for-android-and-iphone.html>.
- [Google, 2011a] Google (2011a). Google Suggestions. [Online; accessed June-2012]. <http://labs.google.com/intl/en/suggestfaq.html>.
- [Google, 2011b] Google (2011b). Sort by subject in Google Images. [Online; accessed June-2014]. <http://googleblog.blogspot.com/2011/05/sort-by-subject-in-google-images.html>.
- [Hearst and Pedersen, 1996] Hearst, M. and Pedersen, J. (1996). Re-examining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 76–84.
- [Kamvar and Baluja, 2006] Kamvar, M. and Baluja, S. (2006). A large scale study of wireless search behavior : Google mobile search. In *Proceedings of 24th Annual SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 701–709.
- [Kamvar and Baluja, 2007] Kamvar, M. and Baluja, S. (2007). Deciphering trends in mobile search. *Computer*, 40(8):58–62.
- [Kamvar and Baluja, 2008] Kamvar, M. and Baluja, S. (2008). Query suggestions for mobile search: Understanding usage patterns. In *Proceedings of 26th Annual SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 1013–1016.
- [Kamvar et al., 2009] Kamvar, M., Kellar, M., Patel, R., and Xu, Y. (2009). Computers and iphones and mobile phones, oh my!: a logs-based comparison of search users on different devices. In *Proceedings of 18th International World Wide Web Conference (WWW)*, pages 801–810.
- [Kong and Allan, 2013] Kong, W. and Allan, J. (2013). Extracting query facets from search results. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 93–102.
- [Krapac et al., 2010] Krapac, J., Moray, A., Verbeek, J., and Jurie, F. (2010). Improving web-image search results using query-relative classifiers. In *IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 1094–1101.
- [Kuroda et al., 2008] Kuroda, M., Sakakihara, M., and Geng, Z. (2008). Acceleration of the em and ecm algorithms using the aitken δ^2 method for log-linear models with partially classified data. *Statistics & Probability Letters*, 78(15):2332–2338.
- [Landauer and Dumais, 1997] Landauer, T. and Dumais, S. (1997). A solution to plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

- [Lau et al., 2013] Lau, H. J., Cook, P., and Baldwin, T. (2013). unimelb: Topic modelling-based word sense induction for web snippet clustering. *Second Joint Conference on Lexical and Computational Semantics (*SEM)*, Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 217–221.
- [Likasa et al., 2003] Likasa, A., N., V., and Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, 36:451–461.
- [Liu et al., 2004] Liu, H., Xie, X., Tang, X., and Ma, W.-Y. (2004). Clustering-based navigation of image search results on mobile devices. In *2004 international conference on Asian Information Retrieval Technology (AIRS)*, pages 325–336.
- [Lloyd, 1982] Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.
- [Martins et al., 2009] Martins, A., Smith, N., Xing, E., Aguiar, P., and Figueiredo, M. (2009). Nonextensive information theoretic kernels on measures. *The Journal of Machine Learning Research*, 10:935–975.
- [McCallum, 2002] McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. <http://mallet.cs.umass.edu>.
- [Métivier et al., 2012] Métivier, J.-P., Boizumault, P., Crémilleux, B., Khiari, M., and Loudni, S. (2012). Constrained clustering using sat. In *Proceedings of the 11th International Conference on Advances in Intelligent Data Analysis (IDA)*, pages 207–218.
- [Meusel et al., 2014] Meusel, R., Vigna, S., Lehmberg, O., and Bizer, C. (2014). Graph structure in the web - revisited. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 427–432.
- [Mihalcea et al., 2006] Mihalcea, R., Corley, C., and Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, pages 775–780.
- [Milligan and Cooper, 1985] Milligan, G. and Cooper, M. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179.
- [Milne and Witten, 2013] Milne, D. and Witten, I. (2013). An open-source toolkit for mining wikipedia. *Journal of Artificial Intelligence*, 194:222–239.
- [Moreno and Dias, 2011] Moreno, J. G. and Dias, G. (2011). Using ephemeral clustering and query logs to organize web image search results on mobile devices. In *Proceedings of International ACM workshop on Interactive Multimedia on Mobile and Portable Devices (IMMPD)*, pages 33–38.

- [Moreno and Dias, 2012] Moreno, J. G. and Dias, G. (2012). Image results exploration using ephemeral clustering. In *Demo Paper Selected by the CNRS for research center representation at World Wide Web Conference (CNRS-WWW)*, pages 1–3.
- [Moreno and Dias, 2013a] Moreno, J. G. and Dias, G. (2013a). Hultech at the ntcir-10 intent-2 task: Discovering user intents through search results clustering. In *Proceedings of the 10th NII Testbeds and Community for Information Access Research Workshop (NTCIR)*, pages 140–146.
- [Moreno and Dias, 2013b] Moreno, J. G. and Dias, G. (2013b). Using text-based web image search results clustering to minimize mobile devices wasted space-interface. In *Proceedings of the 35th European conference on Advances in Information Retrieval (ECIR)*, pages 532–544.
- [Moreno and Dias, 2014a] Moreno, J. G. and Dias, G. (2014a). Easy web search results clustering: When baselines can reach state-of-the-art algorithms. In *Proceedings of 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1–5.
- [Moreno and Dias, 2014b] Moreno, J. G. and Dias, G. (2014b). Pagerank-based word sense induction within web search results clustering. In *Proceedings of the 14th ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–2.
- [Moreno et al., 2013] Moreno, J. G., Dias, G., and Cleuziou, G. (2013). Post-retrieval clustering using third-order similarity measures. In *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 153–158.
- [Moreno et al., 2014] Moreno, J. G., Dias, G., and Cleuziou, G. (2014). Query log driven web search results clustering. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR)*, pages 777–786.
- [Morik et al., 2012] Morik, K., Kaspari, A., Wurst, M., and Skirzynsk, M. (2012). Multi-objective frequent termset clustering. *Knowledge Information Systems*, 30(3):715–738.
- [Müller et al., 2010] Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Radhouani, S., Bakke, B., Kahn, C., and Hersh, W. (2010). Overview of the clef 2009 medical image retrieval track. *Multilingual Information Access Evaluation II. Multimedia Experiments*, 6242:72–84.
- [Navigli and Crisafulli, 2010] Navigli, R. and Crisafulli, G. (2010). Inducing word senses to improve web search result clustering. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 116–126.
- [Navigli and Vannella, 2013] Navigli, R. and Vannella, D. (2013). Semeval-2013 task 11: Word sense induction & disambiguation within an end-user application. In *Proceedings of the International Workshop on Semantic Evaluation (SEMEVAL)*, pages 1–9.

- [Ng et al., 2001] Ng, A., Jordan, M., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In *Proceedings of the 15th Neural Information Processing Systems Conference (NIPS)*, pages 849–856.
- [Osinski et al., 2004] Osinski, S., Stefanowski, J., and Weiss, D. (2004). Lingo: Search results clustering algorithm based on singular value decomposition. In *Intelligent Information Systems Conference (IIPWM)*, pages 369–378.
- [Osinski and Weiss, 2005] Osinski, S. and Weiss, D. (2005). A concept-driven algorithm for clustering search results. *IEEE Intelligent Systems*, 20(3):48–54.
- [Paek et al., 2009] Paek, T., Lee, B., and Thiesson, B. (2009). Designing phrase builder : A mobile real-time query expansion interface. In *Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services (Mobile-HCI)*, pages 7:1–7:10.
- [Pecina, 2005] Pecina, P. (2005). An extensive empirical study of collocation extraction methods. In *Proceedings of the Association for Computational Linguistics Student Research Workshop (ACL)*, pages 13–18.
- [Pecina and Schlesinger, 2006] Pecina, P. and Schlesinger, P. (2006). Combining association measures for collocation extraction. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL)*, pages 651–658.
- [Sakai et al., 2013a] Sakai, T., Dou, Z., Yamamoto, T., Liu, Y., Zhang, M., Kato, M., Song, R., and Iwata, M. (2013a). Summary of the ntcir-10 intent-2 task: Subtopic mining and search result diversification. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 761–764.
- [Sakai et al., 2013b] Sakai, T., Dou, Z., Yamamoto, T., Lui, Y. Zhang, M., and Song, R. (2013b). Overview of the ntcir-10 intent-2 task. In *Proceedings of the Research Infrastructure for Comparative Evaluation of Information Retrieval and Access Technologies (NTCIR)*, pages 94–123.
- [Sakai and Song, 2011] Sakai, T. and Song, R. (2011). Evaluating diversified search results using per-intent graded relevance. In *Proceedings of the 34th international ACM conference on Research and development in Information Retrieval (SIGIR)*, pages 1043–1052.
- [Scaiella et al., 2012] Scaiella, U., Ferragina, P., Marino, A., and Ciaramita, M. (2012). Topical clustering of search results. In *5th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 223–232.

- [Silva et al., 1999] Silva, J., Dias, G., Guilloiré, S., and Lopes, J. (1999). Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. In *Proceedings of 9th Portuguese Conference in Artificial Intelligence (EPIA)*, pages 113–132.
- [Sohn et al., 2008] Sohn, T., Li, K. A., Griswold, W. G., and Hollan, J. D. (2008). A diary study of mobile information needs. In *Proceedings of 26th Annual SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 433–442.
- [Steinbach et al., 2000] Steinbach, M., Karypis, G., and Kumar, V. (2000). A comparison of document clustering techniques. In *Proceedings of the KDD Workshop on Text Mining*, pages 1–2.
- [Timonen, 2013] Timonen, M. (2013). *Term Weighting in Short Documents for Document Categorization, Keyword Extraction and Query Expansion*. PhD thesis, University of Helsinki, Finland.
- [Tolchinsky et al., 2012] Tolchinsky, P., Chiarandini, L., and Jaimes, A. (2012). Prisma: Searching images in parallel. In *Proceedings of the 20th ACM International Conference on Multimedia (MM)*, pages 985–988.
- [Vinh et al., 2009] Vinh, N., Epps, J., and Bailey, J. (2009). Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 1073–1080.
- [Vitale et al., 2012] Vitale, D., Ferragina, P., and Scaiella, U. (2012). Classification of short texts by deploying topical annotations. In *Proceedings of 34th European Conference on Advances in Information Retrieval (ECIR)*, pages 376–387.
- [Wang et al., 2007] Wang, S., Jing, F., He, J., Du, Q., and Zhang, L. (2007). Igroup: Presenting web image search results in semantic clusters. In *Proceedings of 25th Annual SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 587–596.
- [Wang and Zhai, 2007] Wang, X. and Zhai, C. (2007). Learn from web search logs to organize search results. In *Proceedings of 30th Annual International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 87–94.
- [Wang et al., 2004] Wang, X.-J., He, Q.-C., and Li, X. (2004). Grouping web image search result. In *Proceedings of 12th Annual ACM International Conference on Multimedia (MM)*, pages 436–439.
- [Yahoo, 2011] Yahoo (2011). Yahoo Suggestions. [Online; accessed June-2014]. <http://developer.yahoo.com/search/web/V1/relatedSuggestion.html>.

- [Zamir and Etzioni, 1998] Zamir, O. and Etzioni, O. (1998). Web document clustering: A feasibility demonstration. In *Proceedings of 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 46–54.
- [Zeng et al., 2004] Zeng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y., and Ma, J. (2004). Learning to cluster web search results. In *Proceedings of 27th Annual International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 210–217.

Text-Based Ephemeral Clustering for Web Image Retrieval on Mobile Devices.

In this thesis, we present a study about Web image results visualization on mobile devices. Our main findings were inspired by the recent advances in two main research areas - Information Retrieval and Natural Language Processing. In the former, we considered different topics such as search results clustering, Web mobile interfaces, query intent mining, to name but a few. In the latter, we were more focused in collocation measures, high order similarity metrics, etc. Particularly in order to validate our hypothesis, we performed a great deal of different experiments with task specific datasets. Many characteristics are evaluated in the proposed solutions. First, the clustering quality in which classical and recent evaluation metrics are considered. Secondly, the labeling quality of each cluster is evaluated to make sure that all possible query intents are covered. Thirdly and finally, we evaluate the user's effort in exploring the images in a gallery-based interface. An entire chapter is dedicated to each of these three aspects in which the datasets - some of them built to evaluate specific characteristics - are presented.

For the final results, we can take into account two developed algorithms, two datasets and a SRC evaluation tool. From the algorithms, Dual *C*-means is our main product. It can be seen as a generalization of our previously developed algorithm, the *AGK*-means. Both are based in text-based similarity metrics. A new dataset for a complete evaluation of SRC algorithms is developed and presented. Similarly, a new Web image dataset is developed and used together with a new metric to measure the users effort when a set of Web images is explored. Finally, we developed an evaluation tool for the SRC problem, in which we have implemented several classical and recent SRC metrics.

Our conclusions are drawn considering the numerous factors that were discussed in this thesis. However, additional studies could be motivated based in our findings. Some of them are discussed in the end of this study and preliminary analysis suggest that they are directions that have potential.

Keywords: Natural Language Processing; Multimedia Information Retrieval; Ephemeral Clustering; Mobile Devices

Partitionnement Éphémère pour la Recherche d'Images Web en Dispositifs Nomades.

Dans cette thèse, nous présentons une étude sur la visualisation des résultats Web d'images sur les dispositifs nomades. Nos principales conclusions ont été inspirées par les avancées récentes dans deux principaux domaines de recherche – la recherche d'information et le traitement automatique du langage naturel. Tout d'abord, nous avons examiné différents sujets tels que le regroupement des résultats Web, les interfaces mobiles, la fouille des intentions sur une requête, pour n'en nommer que quelques-uns. Ensuite, nous nous sommes concentré sur les mesures d'association lexicale, les métriques de similarité d'ordre élevé, etc. Notamment afin de valider notre hypothèse, nous avons réalisé différentes expériences avec des jeux de données spécifiques de la tâche. De nombreuses caractéristiques sont évaluées dans les solutions proposées. Premièrement, la qualité de regroupement en utilisant à la fois des métriques d'évaluation classiques, mais aussi des métriques plus récentes. Deuxièmement, la qualité de l'étiquetage de chaque groupe de documents est évaluée pour s'assurer au maximum que toutes les intentions des requêtes sont couvertes. Finalement, nous évaluons l'effort de l'utilisateur à explorer les images dans une interface basée sur l'utilisation des galeries présentées sur des dispositifs nomades. Un chapitre entier est consacré à chacun de ces trois aspects dans lesquels les jeux de données - certains d'entre eux construits pour évaluer des caractéristiques spécifiques - sont présentés.

Comme résultats de cette thèse, nous sommes développés : deux algorithmes adaptés aux caractéristiques du problème, deux jeux de données pour les tâches respectives et un outil d'évaluation pour le regroupement des résultats d'une requête (SRC pour les sigles en anglais) . Concernant les algorithmes, Dual *C*-means est notre principal contribution. Il peut être vu comme une généralisation de notre algorithme développé précédemment, l'*AGK*-means. Les deux sont basés sur des mesures d'association lexicale à partir des résultats Web. Un nouveau jeu de données pour l'évaluation complète d'algorithmes SRC est élaboré et présenté. De même, un nouvel ensemble de données sur les images Web est développé et utilisé avec une nouvelle métrique à fin d'évaluer l'effort fait pour les utilisateurs lors qu'ils explorent un ensemble d'images. Enfin, nous avons développé un outil d'évaluation pour le problème SRC, dans lequel nous avons mis en place plusieurs mesures classiques et récentes utilisées en SRC.

Nos conclusions sont tirées compte tenu des nombreux facteurs qui ont été discutés dans cette thèse. Cependant, motivés par nos conclusions, des études supplémentaires pourraient être développés. Celles-ci sont discutées à la fin de ce manuscrit et notre résultats préliminaires suggère que l'association de plusieurs sources d'information améliore déjà la qualité du regroupement.

Mot-clés: Traitement Automatique du Langage; Recherche d'Information Multimédia; Partitionnement Éphémère; Dispositifs Nomades

Discipline: Informatique et applications

Laboratoire: Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen - GREYC CNRS UMR 6072, Sciences 3, Campus 2, Bd Marechal Juin, Université de Caen, 14032 Caen

