Regroupement Automatique de Documents en Classes Événementielles

Aurélien Bossard, Thierry Poibeau

TALN 2008, Avignon, 9–13 juin 2008

Regroupement automatique de documents en classes événementielles

Aurélien Bossard Thierry Poibeau
LIPN - UMR 7030
CNRS - Université Paris 13
F-93430 Villetaneuse, France
{prenom.nom}@lipn.univ-paris13.fr

Résumé. Cet article porte sur le regroupement automatique de documents sur une base événementielle. Après avoir précisé la notion d'événement, nous nous intéressons à la représentation des documents d'un corpus de dépêches, puis à une approche d'apprentissage pour réaliser les regroupements de manière non supervisée fondée sur k-means. Enfin, nous évaluons le système de regroupement de documents sur un corpus de taille réduite et nous discutons de l'évaluation quantitative de ce type de tâche.

Abstract. This paper analyses the problem of automatic document clustering based on events. We first specify the notion of event. Then, we detail the document modelling method and the learning approach for document clustering based on k-means. We finally evaluate our document clustering system on a small corpus and discuss the quantitative evaluation for this kind of task.

Mots-clés : Regroupement de documents, Suivi d'événement.

Keywords: Document clustering, Event tracking.

1 Introduction

La veille est devenue un enjeu majeur pour les entreprises, qu'il s'agisse de veille technique ou scientifique, commerciale ou stratégique. Les « veilleurs » manipulent des masses de données de plus en plus importantes et ont besoin d'aides automatiques afin d'explorer au mieux ces données. Dans cet esprit, de nouvelles perspectives de recherche ont vu le jour afin de faciliter l'accès à un contenu noyé dans un flot d'informations trop important. C'est notamment le cas des tâches de détection et de suivi d'événement (en anglais *topic detection and tracking – TDT*).

La détection et le suivi d'événement consistent à regrouper dans une même classe les documents qui traitent d'un même événement. A titre d'exemple, deux dépêches ayant pour titre « Arrivée en France de Laurent Gbagbo en vue d'une table ronde à Marcoussis » et « Ouverture des négociations entre rebelles et gouvernement ivoirien à Marcoussis » se rapportent à un même événement : « La Table ronde de Marcoussis ». La notion d'événement est cependant une notion vague, qu'il nous appartiendra de préciser au cours de cet article.

On distingue deux cadres applicatifs aux tâches de détection et de suivi d'événement : le cadre « *en ligne* » et le cadre « *hors ligne* ». Dans le premier cas, des documents arrivent les uns à

la suite des autres, et les systèmes mis en œuvre pour traiter la détection d'événement doivent tenir compte de la spécificité des flux d'informations continus. Dans le deuxième cas, les documents à traiter sont déjà tous présents, et la détection d'événement consiste alors à regrouper les documents dans différentes classes qui correspondent chacune à un événement différent.

Nous présentons dans cet article nos travaux sur l'aide à l'analyse de corpus et la détection hors ligne d'événements, c'est-à-dire le regroupement non supervisé de documents selon les événements dont ils traitent. Nous nous fondons pour cela sur une analyse automatique des entités nommées, pour esquisser des liens entre documents. Cette analyse n'étant pas suffisante, nous étudions plusieurs techniques permettant de pondérer les différents types d'entités lors de l'étape de regroupement automatique (*classification*). Ces techniques permettent d'obtenir des « paquets » de documents homogènes du point de vue de l'événement traité, ainsi qu'une visualisation du fonds documentaire dans son ensemble, sous forme d'un réseau social.

Après avoir présenté un état de l'art des technique de détection hors ligne d'événements, nous essayons de mieux caractériser la notion d'événement avant de présenter notre système de détection automatique. Nous détaillons ensuite la façon dont les documents sont caractérisés, puis l'algorithme de classification. Enfin, nous présentons l'évaluation de notre système sur un corpus de dépêches AFP.

2 Etat de l'art

La détection d'événements permet de suivre en direct des flux de dépêches et de les classer en fonction du thème traité. Nous nous intéressons ici à la détection d'événements hors ligne. Ce thème a été moins traité que la détection en ligne mais il est important, au moins dans deux cas de figure bien identifiés :

1. les analystes sont souvent confrontés à des masses de documents traitant de plusieurs thèmes. Avant d'accéder aux documents pertinents, une structuration du fonds documentaire est nécessaire.
2. les systèmes automatiques d'extraction d'information nécessitent des masses de documents homogènes en entrée. Il faut donc les structurer par thème ou par événement avant de passer à la phase d'extraction proprement dite.

Nous poursuivons ces deux buts à la fois, le but de notre application étant *in fine* de produire des synthèses sommaires à partir de masses de documents non structurés. La tâche s'apparente donc à du résumé multi-documents à partir d'un fonds documentaire non homogène en entrée. La visualisation des données permet en outre à l'analyste de contrôler le processus de regroupement de documents en ensembles pertinents. Nous ne nous intéressons ici qu'à l'étape de regroupement des documents.

Plusieurs auteurs ont décrit des systèmes liés à la détection d'événements hors ligne. Il s'agit de (Yang *et al.*, 1999), (Hatzivassiloglou *et al.*, 2000), (Zhiwei Li & Ma, 2005). Les premiers et les seconds utilisent des algorithmes de classification hiérarchique, tandis que les troisièmes utilisent des modèles probabilistes.

(Hatzivassiloglou *et al.*, 2000) se sont posé la question des données à utiliser pour la détection d'événements : vaut-il mieux utiliser la totalité des mots/phrases, exclure des mots qui ne sont

pas catégorisables comme termes uniques (e.g. le Palais, peut être du Luxembourg, de l'Elysée...), ou ne tenir compte que des noms propres ? Les auteurs arrivent à la conclusion que les jeux de données avec lesquels ils obtiennent les meilleurs résultats sont ceux prenant en compte tous les mots sans exception. Ils attribuent cela au fait que les outils d'extraction de termes ou de noms propres qu'ils utilisent ne sont pas assez robustes pour ce type de tâche.

(Yang *et al.*, 1999) proposent une méthode de regroupement de documents où chacun des documents est représenté par une liste de termes pondérés par leur *tf.idf* (cf. *infra*, fig. 2). Sur le programme TDT, en utilisant un algorithme k-means multi-passes, les auteurs arrivent à des résultats de 61 % et 69 % en précision/rappel.

(Zhiwei Li & Ma, 2005) proposent quant à eux une approche probabiliste pour le regroupement de documents en utilisant comme représentation d'un document une matrice composée de quatre vecteurs : les noms de personnes, de lieux, les dates et des mots-clefs. Leur modèle probabiliste appliqué à un extrait du corpus du programme TDT4 produit des résultats de l'ordre de 85 % de précision et 67 % de rappel, en fixant à la main le nombre de classes dans lesquelles ranger les documents. Sur des jeux de données ne séparant pas les entités nommées des mots-clefs, les résultats sont inférieurs de 10 %. Les auteurs l'expliquent par le fait que lorqu'elles ne sont pas distinguées des mots-clefs, les entités nommées se retrouvent noyées dans les données, alors que ce sont les éléments clés pour la construction d'un modèle d'événement.

Les expériences visant à regrouper dynamiquement un flux de documents *en ligne* utilisent globalement les mêmes méthodes, à l'instar de (Binsztok *et al.*, 2004) : il s'agit généralement d'approches probabilistes combinant des sacs de mots et une fenêtre temporelle associée à chaque groupe de documents.

Toutes les approches présentées ici, particulièrement (Hatzivassiloglou *et al.*, 2000), utilisent pour caractériser un document des vocabulaires assez étendus. La taille des données induite par ce type de caractérisation fait chuter les performances et la vitesse des systèmes de classification. Par ailleurs, il a été montré dans (Zhiwei Li & Ma, 2005) que la prise en compte de tout le vocabulaire est moins pertinente que la focalisation sur les seuls éléments clés, notamment les entités nommées. Celles-ci ont par ailleurs un rôle déterminant puisque les fondre dans la masse de données fait chuter les performances.

Enfin, par rapport à des approches comme (Zhiwei Li & Ma, 2005), nous souhaitons élaborer une méthode qui évite d'avoir à préciser à la main *a priori* le nombre de classes visées. Plusieurs solutions existent pour cela, dont l'utilisation de l'indice de Davies-Bouldin (Davies & Bouldin, 1979), une mesure permettant de quantifier la validité d'un clustering. Cet indice correspond au rapport des inerties inter et intra-classes. Parmi les autres méthodes de sélection du nombre de classes, on peut citer l'approche de (Hamerly & Feng, 2006), qui permet d'obtenir une meilleure mesure de la validité du clustering dans des cas difficiles, comme les grandes dimensions. Dans notre cas, l'indice de Davies-Bouldin semble approprié : nous travaillons sur des données de taille modeste. Nous projetons d'examiner ces autres mesures sur des données plus volumineuses.

3 Comment caractériser la notion d'événement ?

Si tous les travaux présentés dans l'état de l'art obtiennent des résultats satisfaisants, aucun n'a tenté de décrire formellement ce qu'est un événement. Donner une définition d'un tel concept

est certes difficile, mais nous avons cependant considéré qu'il était nécessaire de caractériser un événement, ne serait-ce que pour rendre plus objective l'évaluation.

Nous avons exploré différentes définitions d'un événement, notamment d'un point de vue sociologique et d'un point de vue linguistique. D'un point de vue sociologique tout d'abord, l'événement prend autant de définitions que de champs disciplinaires dans lesquels il est considéré (Prestini-Christophe, 2006). Il existe cependant des points communs à toutes ces définitions :

– un événement est un fait inattendu et correspond à une rupture,
– un fait devient événement en fonction du monde dans lequel il advient : l'événement est subjectif.

D'un point de vue linguistique, Pustejosky a élaboré une théorie (Pustejovsky, 2000) dans laquelle il fonde sa définition de l'événement sur le repérage de structures prédicat/arguments et sur la notion de changement. L'événement est identifiable grâce à un *prédicat d'événement* (i.e. un verbe qui implique un changement) et une structure argumentale équivalente (c'est-à-dire une identité référentielle des arguments du prédicat, même si les formes de surface employées sont différentes). Une implémentation de cette théorie nécessiterait des connaissances sémantiques très fines sur les verbes, mais également des connaissances sur les différents arguments possibles.

Etant donné le nombre de définitions et le peu de formalisation du concept d'événement, il est sûrement plus judicieux de partir de nos besoins afin de définir le concept d'événement dans le cadre de notre travail. Notre tâche consiste à regrouper les documents d'actualité (des dépêches) qui traitent du même événement, afin de réaliser une synthèse automatique des groupes créés. On part du principe qu'une dépêche traite d'un événement unique (ceci est généralement vérifié, sauf pour certaines dépêches qui retracent tous les faits marquants d'une journée, ou le déroulement d'une succession de faits plus ou moins liés entre eux). Les dépêches sont en outre rédigées en forme de pyramide inversée : le premier paragraphe doit normalement contenir l'information sur l'événement brut, les détails, commentaires et opinions étant normalement détaillés dans la suite du document.

Il s'agit alors, afin de réaliser notre tâche, de trouver les traits communs entre les dépêches qui traitent d'un même événement. Un événement consiste en une action réalisée par une ou plusieurs personnes, à une certaine date, dans un certain lieu. Il est donc assez intuitif d'utiliser ces marqueurs afin de reconnaître que deux dépêches traitent du même sujet. Nous retrouvons ici les éléments mis en évidence par (Pustejovsky, 2000) concernant les arguments du prédicat. On peut par ailleurs utiliser des mots clefs, tels que « procès », « élections », « bombardement », afin de départager des dépêches partageant des entités nommées mais dont le lecteur a l'intuition qu'ils se rapportent malgré tout à des événements différents.

Reste à définir la portée d'un événement. En d'autres termes, quelle fenêtre temporelle doit-on adopter pour exclure d'un groupe un document qui traite du même événement qu'un document précédent ? Contrairement à (Binsztok *et al.*, 2004), nous considérons la notion de portée indépendante de toute notion temporelle. Un fait directement lié à un événement initial peut se produire longtemps après le fait initial, comme la reconnaissance de l'innocence d'un homme vingt ans après sa condamnation.

Afin d'éviter toute ambiguïté, nous appellerons dorénavant l'ensemble des dépêches regroupées ensemble, parlant de faits directement liés entre eux, le « sujet », et le fait initial, celui qui aura conduit à cette succession de faits, l'« événement ». Les regroupements que nous chercherons à effectuer seront donc des regroupements par sujet, et non des regroupements par événement.

Cette notion de sujet se rapproche de la notion d'« activité » telle que définie dans TDT4, par opposition à la notion d'« événement », toujours dans TDT4.

Malgré cette tentative d'« objectivisation » des notions d'événement et de sujet, ces notions restent largement subjectives. A titre d'exemple, prenons deux documents : l'un parlant des élections anticipées en Côte d'Ivoire consécutives aux accords de Marcoussis, l'autre parlant de la table ronde de Marcoussis. Faut-il créer deux sujets pour ces deux dépêches – l'un concernant la réunion de Marcoussis, l'autre l'application des accords décidés à Marcoussis, voire les élection anticipées – ou les réunir au sein d'un même sujet concernant les accords de Marcoussis, de leur négociation à leur application ? Il y a là une part de subjectivité qui ne peut être complètement évitée.

4 Modèle de description des documents

Afin de regrouper les documents par sujet, mais également afin d'obtenir une représentation graphique qui permette à un utilisateur de prendre rapidement connaissance du contenu d'un fonds documentaire, celui-ci est représenté à la manière d'un réseau social (fig 1). Chaque document a des liens avec les autres documents du corpus. Le poids des liens entre deux documents est calculé selon les entités nommées qu'ils partagent.

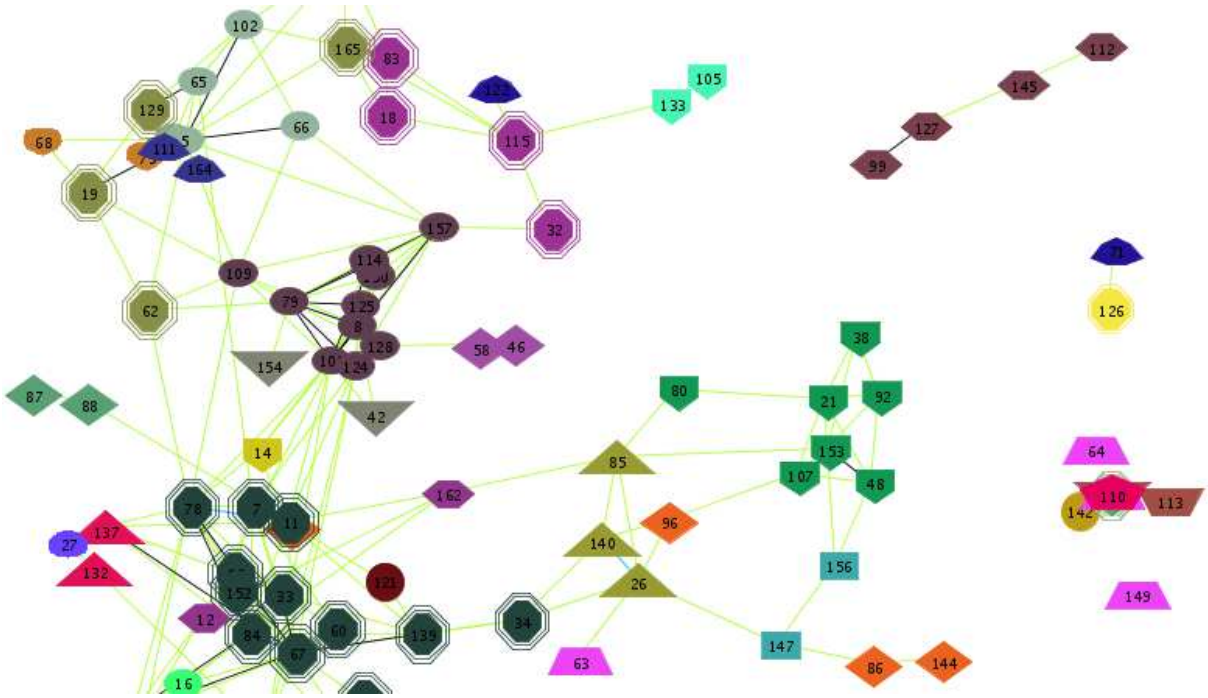

FIG. 1 – Un corpus vu comme un réseau social : les noeuds sont les documents, leurs forme et couleur dénotent leur appartenance à une classe, et la couleur des liens est fonction de leur poids

Nous avons utilisé, pour calculer la similarité de contenu entre deux documents, un indice Jaccard pondéré. L'indice Jaccard est une métrique utilisée en statistique pour comparer la similarité de deux ensembles, fondée sur le rapport entre la cardinalité de l'intersection et la cardinalité de l'union des ensembles.

Chaque ensemble est formé des entités nommées contenues dans chaque document. Afin d'affiner l'importance relative des différentes entités, il est nécessaire de les pondérer. En effet, les corpus privilégient souvent certains thèmes : certaines entités centrales apparaissent alors de manière relativement uniformes et ont de ce fait un pouvoir discriminant très faible. Leur participation au calcul de similarité doit être en conséquence. Pour cela, nous avons testé deux mesures : la mesure `idf` et la mesure du `tf.idf` (Salton & Buckley, 1988) (nous utilisons le mode de calcul classique de ces deux mesures, cf. figure 2). Cette pondération permet en outre de prendre en compte la fréquence d'un terme. Il y a en effet une très forte probabilité que deux documents qui contiennent les mêmes entités nommées avec des fréquences proches fassent partie d'un même sujet, et la mesure jaccard pondérée par le `tf-idf` permet de rendre compte de ce fait.

Cette pondération selon le `tf.idf` comme réalisée dans (Yang *et al.*, 1999) ne suffit cependant pas : certains types d'entités nommées ont un rôle plus important que d'autres dans la description d'un événement. En effet, les types d'entités les plus instables au sein d'un même sujet sont les types « Personnes » et « Organisations ». Si les références aux dates et aux lieux restent globalement inchangées dans le suivi d'un sujet, les intervenants sont en revanche multiples et variables. Il faut donc distinguer les entités suivant leur type afin d'éviter la création de sujets distincts correspondant à chaque intervenant. Nous avons alors affecté un poids à chaque type d'entités nommées, en privilégiant les lieux et les dates par rapport aux noms de personnes et d'organisations. Le calcul de la similarité entre documents est celui de la figure 3, le `tf.idf` d'une entité pouvant être remplacé par l'`idf`.

5 Classification

Nous réalisons un regroupement des documents en utilisant l'algorithme k-means (Forgy, 1965). Cet algorithme réalise une classification en k classes, en minimisant la variance intra-classe. L'algorithme se déroule en 4 étapes :

1. Choisir aléatoirement k objets qui seront les centres de k classes ;
2. Parcourir tous les objets, et les affecter ou les réaffecter à la classe qui minimise la distance entre l'objet et le centre de la classe ;
3. Calculer les barycentres de chaque classe, ils deviennent les nouveaux centres ;

Calcul du tf d'un terme t_i dans le document d_j :
$\text{tf}_{t_i,d_j} = \frac{|t_i|_j}{\sum_{k=0}^{n} |t_k|_{d_j}}$

Calcul de l'`idf` d'un terme t_i au sein d'un corpus D de documents d_j :
$\text{idf}_{t_i} = \log \frac{(|D|)}{\{|d_j : t_i \in d_j\}|}$

Calcul du `tf.idf` du terme t_i pour un document d_j :
$\text{tf.idf}_{t_i,d_j} = tf_{t_i,d_j} \times idf_{t_i}$

FIG. 2 – Calcul du `tf.idf`

- $S(i,j) = \frac{N_{11}(i,j)}{(N_{11}(i,j)+N_{10}(i,j)+N_{01}(i,j))}$
- $N_{11}(i,j) = \sum_{k=0}^{n} poids(e_k) \times tf.idf(e_k)$, où les ek sont les entités nommées présentes dans i et j.
- $N_{10}(i,j) = \sum_{k=0}^{n} poids(e_k) \times tf.idf(e_k)$, où les ek sont les EN présentes seulement dans i.
- $N_{01}(i,j) = \sum_{k=0}^{n} poids(e_k) \times tf.idf(e_k)$, où les ek sont les EN présentes seulement dans j.

FIG. 3 – Calcul de la mesure de similarité entre documents

4. Répéter les étapes 2 et 3 jusqu'à convergence. La convergence est atteinte lorsque les classes deviennent stables.

L'algorithme k-means prend comme paramètre le nombre de classes (k). Dans le cadre du regroupement non-supervisé de documents, il est nécessaire de laisser ce paramètre libre, et de trouver le meilleur k possible. Afin d'automatiser la recherche du meilleur paramètre k, l'algorithme k-means est appliqué n fois en incrémentant k à chaque fois. Finalement, le paramètre k minimisant l'indice de Davies-Bouldin (Davies & Bouldin, 1979) est retenu (cet indice permet de calculer la validité de la classification, en mesurant la cohérence des regroupements ; les regroupements qui minimisent la distance entre objets de la même classe et maximisent la distance entre objets de classes différentes ont le meilleur indice).

6 Evaluation

Nous avons choisi d'évaluer le travail de classification par un ensemble de mesures quantitatives. Une évaluation qualitative des résultats reste à mener.

6.1 Description du cadre applicatif

Cette recherche s'inscrit dans le cadre du projet Infomagic, du pôle de compétitivité Cap Digital[1]. Ce cadre nous permet d'avoir accès à des besoins opérationnels précis et à des corpus variés.

Un de ces corpus largement mis à contribution dans le cadre d'Infomagic est un ensemble de dépêches AFP portant sur la Côte d'Ivoire. Le corpus compte 15000 documents. L'annotation des documents avec les entités nommées nous a été fourni par la société Arisem (partenaire du projet Infomagic). Nous avons travaillé sur un extrait du corpus comptant 200 dépêches, afin de pouvoir évaluer au mieux la classification. Ainsi, nous avons fait le choix de réaliser une double annotation, afin de pouvoir comparer notre approche à deux résultats obtenus manuellement par des personnes différentes. Le corpus compte 1030 entités nommées différentes. Nous avons

[1]Cap Digital porte sur l'indexation multimedia. Cette recherche s'insère dans le cadre de l'« axe texte » du projet, qui regroupe des entreprises et des laboratoires de recherche en traitement des langues, ainsi que des industriels ayant des besoins spécifiques qui servent de cadres d'application communs.

établi un corpus de référence en classant les dépêches par sujet, et obtenu 44 classes avec une répartition très inégale.

Nous avons voulu évaluer toutes les étapes de la classification, afin d'identifier les faiblesses de la méthode utilisée. Dans un premier temps, nous avons évalué la qualité de la classification en examinant les résultats de l'algorithme k-means en faisant varier k de 2 à 100, ce qui correspond à diviser le corpus en un nombre de classes allant de 2 à 100. Dans un second temps, nous avons évalué le résultat « optimal » de l'apprentissage, qui correspond à la classification qui minimise l'indice de Davies-Bouldin en le comparant au meilleur résultat obtenu sur tous les lancements de k-means, qualitativement parlant.

6.2 Evaluation globale

Nous avons évalué la pertinence de la classification par deux méthodes :

– les micro-moyennes ;
– les macro-moyennes.

Ces deux mesures donnent des résultats très différents : le résultat de la macro-moyenne tient plus compte des catégories ayant peu de documents pertinents, tandis que le résultat des micro-moyennes fait plus ressortir les résultats sur les plus grosses classes. Ces deux mesures sont donc nécessaires pour la bonne interprétation des résultats. La méthode des macro-moyennes consiste à comparer chaque classe C_i du corpus étiqueté automatiquement à la classe du corpus de référence majoritaire dans C_i. Les deux classifications comparées doivent avoir le même nombre de classes. Nous avons donc fixé k au nombre de classes du corpus de référence. Une moyenne est effectuée sur les mesures de chaque classe, avec un poids égal pour chaque classe. La méthode des micro-moyennes consiste à fusionner les tables de contingence de toutes les classes et à calculer les mesures sur la table fusionnée. Nous avons utilisé les mesures classiques de précision, rappel et F-mesure. La F-Mesure est la moyenne harmonique de la précision et du rappel, et favorise les systèmes qui ont des mesures de rappel et de précision proches.

	Précision	Rappel	F-Mesure
Macro-Moyenne	61.4%	65.3%	63.2%
Micro-Moyennes	52.3%	55.6%	53.9%

FIG. 4 – Résultats de l'évaluation avec le nombre de classes fixé

Il est intéressant de noter que deux annotateurs humains ont obtenu sur ce même corpus, des résultats dont la F-Mesure est à 44%. Les deux annotations différentes se défendent, les résultats de celles-ci dépendant fortement du choix de granularité utilisé par les annotateurs dans les sujets, et du choix de raccorder certains événements à un sujet ou de créer un novueau sujet pour ceux-ci. Ceci pose le problème de l'évaluation d'une classification par des méthodes quantitatives.

6.3 Sélection du k par l'indice de Davies-Bouldin

Nous avons évalué la sélection du k selon les deux points suivants :
– la différence entre le nombre de classes choisi automatiquement et le nombre de classes du corpus de référence ;

– le rapport entre l'indice de Davies-Bouldin et une mesure d'évaluation de la classification.
En moyenne sur 15 lancements de l'algorithme (donc 15 configurations de départ différentes), nous avons trouvé 37,4 classes contre 44 classes dans le corpus de référence. Le même corpus annoté par un deuxième annotateur contient quant à lui 65 classes, l'annotateur ayant fait des choix différents concernant les sujets existants. On constate donc une erreur dans le choix automatique du nombre de classes pouvant aller de 16% à 57%. Cette mesure n'est donc pas asez significative pour évaluer la qualité du choix du nombre de classes. Nous avons évalué la pertinence de l'utilisation de l'indice de Davies-Bouldin ; la figure 5 montre l'évolution de l'indice pour tous les k ainsi qu'une mesure d'évaluation du clustering, qui consiste à considérer tous les documents liés au sein d'un même cluster, et à effectuer les calculs de précision/rappel sur la présence/absence des liens du corpus de référence dans le corpus clusterisé automatiquement.

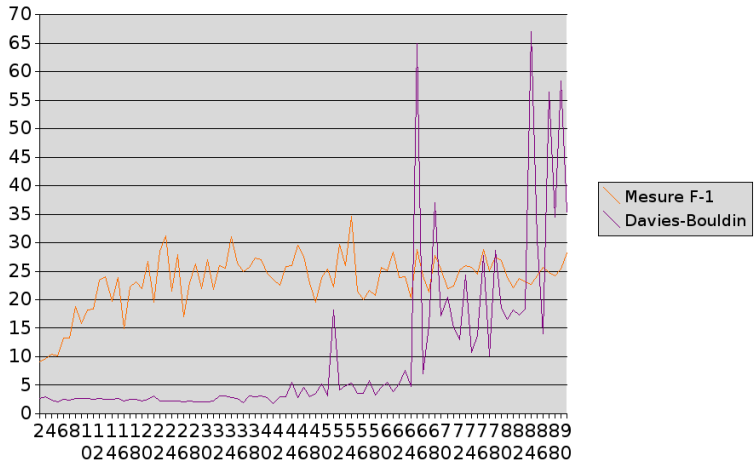

FIG. 5 – Indice de Davies-Bouldin et Choix du k

On constate la non-corrélation de l'indice de Davies-Bouldin et le choix d'un nombre de classes qui maximise la F-mesure. Ceci est en grande partie dû au fait que le découpage du corpus de référence n'est pas homogène : une des classes contient à elle seule 1/6 des documents du corpus. Or, le choix d'un k par Davies-Bouldin et la classification des documents par k-means seront optimaux dans le cas où les classes ont toutes approximativement le même nombre d'objets.

7 Conclusion

Nous avons présenté dans cet article un système de regroupement de dépêches, fondé sur les entités nommées partagées entre documents. Les résultats obtenus suite à la classification automatique sont supérieurs à ceux obtenus par des annotateurs humains.

L'évaluation quantitative devrait à l'avenir être complétée par une évaluation qualitative. En effet, les systèmes d'évaluation quantitative ne permettent pas de valider l'utilisabilité des résultats : la séparation d'une classe en deux par un algorithme peut faire chuter le rappel de 50%. Les résultats obtenus automatiquement sont toutefois intéressants et peuvent faire ressortir d'autres regroupements que ceux choisis par les auteurs du corpus de référence, le regroupement de documents étant très subjectif.

L'algorithme utilisé pour la classification, k-means, n'est pas exempt de défauts : moins efficace sur des données à regrouper dans des classes non-homogènes, il devient également moins performant sur des corpus de grande dimension. Pour passer à l'échelle supérieure, il nous faudra donc explorer d'autres méthodes de classification, comme les SVM et les cartes de Kohonen non supervisées.

Remerciements

Ces recherches sont en partie financées à travers le projet Infomagic de Pôle de compétitivité Cap Digital. Nous remercions en particulier la société Arisem qui nous a fourni l'annotation des entités nommées.

Références

BINSZTOK H., ARTIÈRES T. & GALLINARI P. (2004). Un modèle probabiliste de détection en ligne de nouvel evénement. In *Reconnaissance des Formes et Intelligence Artificielle (RFIA 2004)*, Toulouse, France.

DAVIES D. L. & BOULDIN D. W. (1979). A cluster separation measure. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, p. 224–227.

FORGY E. (1965). Cluster analysis of multivariate data : Efficiency vs. interpretability of classifications. *Biometrics*, p. 21–768.

HAMERLY G. & FENG Y. (2006). Pg-means : learning the number of clusters in data. In *The Twentieth Annual Conference on Neural Information processing systems*, Vancouver, Canada.

HATZIVASSILOGLOU V., GRAVANO L. & MAGANTI A. (2000). An investigation of linguistic features and clustering algorithms for topical document clustering. In *SIGIR 2000 : Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 224–231, Athens, Greece : ACM.

PRESTINI-CHRISTOPHE M. (2006). La notion d'evénement dans différents champs disciplinaires. *Pensée Plurielle*, **13**, 21–29.

PUSTEJOVSKY J. (2000). Events and the semantics of opposition. In C. TENNY & J. PUSTEJOVSKY, Eds., *Events as Grammatical Objects*, chapter 13, p. 445–482. CSLI Publications.

SALTON G. & BUCKLEY C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management : an International Journal*, **24**, 513–523.

YANG Y., CARBONEL J. G., BROWN R. D., PIERCE T. & BRIAN T. ARCHIBALD X. L. (1999). Learning approaches for detecting and tracking news events. In *IEEE Intelligent Systems*, p. 32–43, Cambridge, Massachusetts.

ZHIWEI LI, BIN WANG M. L. & MA W.-Y. (2005). A probabilistic model for retrospective news event detection. In *SIGIR 2005 : Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil : ACM.