APPROCHES DE CLASSIFICATION POUR LE FILTRAGE DE DOCUMENTS IMPORTANTS AU SUJET D'UNE ENTITÉ NOMMÉE

Ludovic Bonnefoy, Vincent Bouvier, Patrice Bellot

Lavoisier | « Document numérique »

2014/1 Vol. 17 | pages 9 à 36

ISSN 1279-5127
ISBN 9782746246546

Article disponible en ligne à l'adresse :

https://www.cairn.info/revue-document-numerique-2014-1-page-9.htm

Approches de classification pour le filtrage de documents importants au sujet d'une entité nommée

Ludovic Bonnefoy [1], **Vincent Bouvier** [2,3], **Patrice Bellot** [3]

1. *Université d'Avignon et des Pays de Vaucluse, LIA*
   Agroparc - BP 91228 - 339, chemin des Meinajariès
   84911 Avignon Cedex 9, France
   firstname.lastname@univ-avignon.fr

2. *Kware*
   565 Rue Marcelin Berthelot
   13851 Aix-en-Provence Cedex 3, France
   firstname.lastname@kware.fr

3. *Aix-Marseille Université CNRS, LSIS UMR 7296*
   Av. Escadrille Normandie Niemen
   13397 Marseille Cedex 20, France
   firstname.lastname@univ-amu.fr

RÉSUMÉ. *Nous souhaitons filtrer un flux de documents web selon qu'ils mentionnent ou non une entité donnée, tout en mesurant l'importance de l'information présente concernant cette entité. Notre approche repose sur l'utilisation de classifieurs prenant en compte des indices comme la fréquence des mentions de l'entité au fil du temps et dans les documents, leurs positions ou encore la présence d'entités liées connues. Notre approche a été évaluée via les tâches "Knowledge Base Acceleration" de TREC 2012 et 2013, et classée parmi les plus performantes.*

ABSTRACT. *Our aim is to filter a stream of Web documents according to whether they refer or not an entity, while estimating the importance of the information contained about this entity. Our approach relies on the use of classifiers taking into account features such as the frequency of the entity over time and in the documents, their positions and the presence of known related entities. Our approach was evaluated during "Knowledge Base Acceleration" tracks of TREC 2012 and 2013 and has been ranked among the best ones.*

MOTS-CLÉS : *filtrage, entité nommée, TREC KBA, forêt d'arbres décisionnels, classification.*

KEYWORDS: *Filtering, named entity, TREC KBA, random forest, classification.*

DOI:10.3166/DN.17.1.9-36

1. Introduction

Les entités nommées sont au cœur de nombreux travaux dans le domaine de la recherche d'information et du traitement automatique des langues. Cet intérêt a été impulsé et maintenu vivace grâce à de multiples campagnes d'évaluations : MUC (*Named Entity task* [1]), ACE (*Entity Mention Detection task* (Doddington *et al*., 2004), TREC (avec la tâche *Question Answering* (Voorhees, 1999)) etc.

Les premières méthodes non supervisées de recherche d'entités nommées étaient basées sur des ensembles de patrons d'extraction (Nadeau, Sekine, 2007) et, aujourd'hui encore, il est conseillé de procéder de la sorte si un corpus d'entraînement n'est pas disponible pour les types souhaités (Sekine, Nobata, 2004). Avec l'arrivée des premiers corpus d'apprentissage pour quelques types d'entités (personne, lieu, organisation et date), des approches supervisées sont apparues recourant aux modèles de Markov cachés (Bikel *et al*., 1997), aux arbres de décision (Sekine, 1998) ou encore aux SVM (Asahara, Matsumoto, 2003) et CRF (McCallum, 2003). Des méthodes faiblement (ou semi-)supervisées journalName ont aussi été étudiées en exploitant différents critères comme les relations syntaxiques (Cucchiarelli, Velardi, 2001) ou synonymiques (Pasca *et al*., 2006).

Aujourd'hui, le domaine a atteint une certaine maturité mais les performances stagnent quelque peu même s'il reste d'importants progrès à réaliser notamment pour gérer tout type d'entités et pas seulement les types de haut niveau comme les noms de personnes et les noms de lieux (on notera cependant l'existence de travaux traitant de la hiérarchisation des entités nommées comme ceux proposés dans le cadre de QUAERO (Galibert *et al*., 2011)). Les travaux sur le sujet, bien que nombreux, se concentrent désormais sur des sous-problèmes présentant des caractéristiques et difficultés très spécifiques telles que la reconnaissance des entités dans le domaine biomédical (très difficiles à segmenter) (Atkinson, Bull, 2012) ou dans les tweets (très peu de contexte immédiatement accessible) (Liu *et al*., 2012).

De nouvelles applications ont émergé. Parmi elles, deux nous intéressent particulièrement : la résolution des co-références des entités nommées au sein d'un document ou entre plusieurs documents ainsi que la tâche *Entity Linking* (EL) de TAC (*Text Analysis Conference*). Ces deux tâches répondent au problème soulevé par l'ambiguïté des noms des entités : une même entité peut être désignée par plusieurs mots différents et, à l'inverse, un même mot peut désigner plusieurs entités.

La première tâche correspond à lier entre elles, dans un document ou dans une collection, différentes expressions ayant la même référence dans un contexte unique : par exemple "Elvis" et "The King" (ce dernier terme peut référer à Elvis Presley mais aussi, bien sûr, à d'autres personnes dans un contexte différent). Les approches existantes se distinguent d'une part par leur niveau de supervision mais aussi par le niveau auquel sont estimés les différents paramètres pour la prise de décision : de nombreuses

1. http://cs.nyu.edu/faculty/grishman/NEtask20.book_1.html

approches résolvent les co-références en analysant les expressions par paires tandis que d'autres choisissent un angle plus large et travaillent sur l'ensemble des mentions à la fois. Dans tous les cas, les critères utilisés sont des éléments classiques : critères lexicaux (par exemple le recouvrement lexical entre les deux occurrences), sémantiques (mise en relation via des ressources terminologiques telles que Wordnet) ou encore syntaxiques (structures syntaxiques dans lesquelles apparaissent les multiples occurrences). Un état de l'art du domaine est proposé par (Clark, Gonzalez-Brenes, 2008).

Une seconde tâche, *Entity Linking*, se propose de désambiguïser les mentions d'entités rencontrées dans un document en les liant à des entités mentionnées dans une base de connaissances. Cette tâche a vu le jour grâce à l'émergence de bases de connaissances et d'encyclopédies de plus en plus complètes telles que Wikipédia [2] ou Freebase [3]. Outre la désambiguïsation en elle-même, cette tâche présente d'autres intérêts tant du côté des interfaces utilisateur (par l'ajout de liens explicites des occurrences vers les entités elles-mêmes) que du côté des systèmes automatiques comme première étape vers une population automatique des bases de connaissances.

Cet axe de recherche est soutenu par la campagne d'évaluation *Text Analysis Conference* (TAC) [4] avec la tâche *Entity Linking* au sein de la piste *Knowledge Base Population* (KBP) qui met à disposition plusieurs milliers de *topics* (couples "mention d'entité et document de presse associé") ainsi que les jugements de pertinence correspondants, une base de connaissances (créée à partir de Wikipédia) et un corpus fort de plusieurs centaines de milliers de documents journalistiques, de pages web et de blogs. Les approches implémentées par les participants s'apparentent en général à celles des méthodes de résolution des co-références, combinées ou non à l'utilisation de techniques de retour de pertinence et d'expansion de requêtes (Artiles *et al*., 2011). La résolution de cette tâche nécessite de déterminer si l'occurrence d'une entité correspond ou non à une entité déjà présente dans la base de connaissances. Pour cela, des mesures de similarité entre les entrées de la base et le contexte de l'occurrence ainsi que des mesures surfaciques entre les mots sont utilisées. Le compte-rendu de la tâche KBP à TAC 2011 (Ji *et al*., 2011) présente un nombre important d'approches ainsi qu'une évaluation approfondie des performances obtenues. Sans surprise, les approches les plus performantes sont majoritairement supervisées et permettent d'estimer automatiquement le poids de nombreux critères qu'ils soient syntaxiques ou lexicaux.

Il est possible de considérer cette tâche de façon inversée en partant d'entités dans une base de connaissances et en cherchant, dans un flux de documents datés, ceux qui les mentionnent de la façon la plus pertinente ou, autrement dit, qui apportent le plus d'informations nouvelles sur les entités. Les premiers travaux se sont orientés vers l'exploitation des réseaux sociaux, en particulier Twitter, avec pour objectif de

2. http://www.wikipedia.org/
3. http://www.freebase.com/
4. http://www.nist.gov/tac/2012/KBP/index.html

pouvoir suivre l'évolution en direct d'évènements nommés, comme les catastrophes naturelles (Lee, 2012). Cette contrainte de réactivité et de prise de décision en temps réel apporte de nouveaux défis. En effet, de nombreuses approches utilisées pour la désambiguïsation des entités nommées ont recours à des ressources externes ou à des traitements coûteux (telle qu'une analyse syntaxique ou une résolution des références anaphoriques) qui ne sont pas applicables facilement en temps réel. De plus, au cours du temps, les entités évoluent, aussi bien du point de vue des informations qui leur sont associées que de la façon de les dénommer. Si leur représentation initiale n'est pas mise à jour, celle-ci peut devenir obsolète et conduire à ignorer de nouvelles données importantes. L'utilisation de sources informationnelles telles que Twitter est problématique en particulier à cause de la taille des messages qui incite à ne travailler qu'avec des entités très populaires, présentes dans de gros volumes de tweets (Davis *et al*., 2012).

En 2012, une nouvelle piste au sein des conférences TREC, nommée *Knowledge Base Acceleration* (KBA) [5] a été créée. Elle correspond à la tâche suivante : dans une importante base de connaissances comme Wikipédia, il y a plus d'entités qu'il n'y a de contributeurs pour la mettre à jour. Cela engendre un temps de latence médian de 356 jours avant qu'un élément nouveau, en rapport avec une entité donnée, ne soit reporté sur sa page (Frank *et al*., 2012). Ne pourrait-on pas raccourcir ce délai en soumettant à des contributeurs en charge d'une entrée de la base tous les nouveaux documents mentionnant l'entité tout en ordonnant ceux-ci selon une estimation automatique de leur importance ? La tâche KBA de TREC 2012 consiste ainsi à trouver, pour 29 entités de Wikipédia choisies pour leur ambiguïté, tout document les mentionnant au sein d'une large collection et à attribuer à ces derniers une classe d'importance : non pertinent (bien que le document mentionne l'entité, aucune information importante ou précise la concernant n'est présente), pertinent ou vital. Un document est considéré comme vital s'il apporte une information majeure sur l'entité qui doit absolument figurer dans l'entrée correspondante (ici, sa page Wikipédia). Par exemple, pour l'entité *Barack Obama*, un document mentionnant une visite diplomatique est pertinent mais non vital tandis qu'un document annonçant sa réélection l'est. Notons que la question de la nouveauté de l'information avait été mise de côté pour l'édition 2012 de la tâche TREC KBA.

Cette tâche a de fortes similarités avec la tâche *Filtering* organisée à TREC à la fin des années 1990 (S. Robertson, Soboroff, 2002). Celle-ci était définie comme suit : à partir d'une requête utilisateur ainsi que d'un ensemble de documents jugés pertinents, déterminer, pour chaque nouveau document apparaissant dans un flux, s'il répond ou non aux attentes de l'utilisateur. Nombre d'approches eurent recours à des techniques traditionnelles en recherche d'information (par ex. Okapi (S. Robertson *et al*., 2002), Rocchio...) pour associer un score aux documents. Un seuil était ensuite calculé à partir du jeu de documents donné et d'un corpus d'apprentissage. Pour la majorité des autres approches performantes, des SVMs furent utilisés (Cancedda *et al*., 2002), prin-

5. http://trec-kba.org/

cipalement selon des critères lexicaux (comme le compte des n-grammes (Mihalcea, 2002)). Cependant, des différences existent qui rendent difficiles la réutilisation de ces approches pour KBA : la taille de la collection était cent fois plus petite, elle ne contenait que des documents journalistiques (le corpus de KBA contient aussi des documents provenant de blogs et des pages web), l'unité de temps considérée était la journée (contre l'heure pour KBA) – elle permettait d'avoir plus de recul sur un évènement – et, enfin, à chaque prise de décision les systèmes étaient informés de la classe associée au document par les annotateurs (cela permettait de ré-estimer dynamiquement les modèles et de réduire la divergence du modèle initial au fil du temps).

Malgré (ou à cause de) toutes ces difficultés supplémentaires, les meilleurs systèmes de KBA 2012 restent *simples*. Le système qui a obtenu les meilleurs résultats (Kjersten, McNamee, 2012) utilise une représentation vectorielle du document avec pour composantes les mots et entités nommées présents. Les valeurs de ce vecteur sont binaires (absence ou présence du mot ou de l'entité dans le document). Un classifieur de type *séparateurs à vaste marge (SVM)* est ensuite entraîné en utilisant les exemples positifs et négatifs associés à l'entité donnée. 100 000 documents sont pris au hasard dans le corpus comme exemples négatifs supplémentaires (le même jeu de 100 000 documents est utilisé pour chaque entité). Un classifieur est entrainé par entité à suivre. La seconde approche (Liu, Fang, 2012) repose sur l'analyse des entités liées à l'entité étudiée. Ces entités correspondent au nom des articles de Wikipédia pointés par l'article correspondant à l'entité étudiée. Le score final d'un document du flux contenant l'entité est fonction du nombre d'entités liées qu'il contient et de leur poids (nombre d'occurrences).

Plutôt que d'essayer de construire un modèle de document pertinent par entité, à l'instar des méthodes habituelles, nous souhaitons déterminer si les documents contenant des informations importantes au sujet d'une entité ont des caractéristiques particulières, et ce indépendamment de l'entité étudiée, liées par exemples à la structure du document ou encore à sa période d'apparition. Notre approche, obtenant actuellement les meilleurs résultats publiés sur cette tâche, est basée sur l'utilisation de classifieurs pour déterminer le degré d'intérêt d'un document. Les éléments étudiés appartiennent à trois catégories : prise en compte d'informations temporelles, de caractéristiques des occurrences des mots-entités dans le document et de la présence d'entités liées.

Dans cet article, notre approche est présentée dans une première partie (choix du classifieur et des critères utilisés), puis nous analysons et commentons les résultats obtenus dans le cadre de TREC KBA 2012. Dans une dernière partie, nous évoquons les travaux en cours qui ont été partiellement évalués dans le cadre de TREC KBA 2013.

2. Détection de documents centrés sur une entité dans un flux de documents

Considérons un flux de documents de types variés (journalistique, blog, forum, page web, tweets etc.) et une entité donnée e. Nous souhaitons être en mesure de

déterminer si un nouveau document d dans ce flux fait référence à l'entité e et mesurer l'importance de l'information qui y est contenue au regard de cette entité.

Comme cela a déjà été dit, nous ne souhaitons pas construire un modèle de document pertinent par entité mais déterminer si les documents contenant des informations au sujet d'une entité ont des caractéristiques particulières qui permettent ou non de décider de son importance, et ce indépendamment de l'entité étudiée. Ainsi, nous proposons et évaluons dans ce qui suit un ensemble de critères quantitatifs.

Le problème peut être décomposé en deux (les caractéristiques des documents pouvant servir à résoudre chaque sous-problème ne sont peut-être pas les mêmes) :

– déterminer si le document apporte des informations sur l'entité e. Cette première étape a pour rôle le filtrage des documents qui, bien que contenant le "nom" de l'entité, ne mentionnent pas e : le document peut alors faire référence à un homonyme ou alors le nom de e est composé de noms communs (par exemple le groupe de musique *Basic Element*). De plus, cette étape doit aussi filtrer les documents dans lesquels l'entité e est certes citée mais dans lesquels aucune information à son sujet n'est donnée ou cette information n'est pas pertinente pour e ("Hier soir, X, le fils de e, a donné un concert privé [...]") ;

– déterminer l'importance des informations que contient le document d au regard de l'entité e. Nous envisagerons deux cas : soit le document est pertinent, soit il est vital. Un document dit vital pour e contiendra des informations jugées suffisamment importantes pour justifier une mise à jour immédiate de l'entrée de e dans une base de connaissances.

De nombreux critères ont déjà été proposés pour caractériser des documents. Le lecteur pourra se référer à des états de l'art tels que ceux proposés par (Sebastiani, 2002) pour les documents en général ou (Qi, Davison, 2009) pour les pages web. Ces critères varient évidemment en fonction des applications et vont de l'analyse du contenu à celle de la structure ou de l'aspect visuel du document, en passant par l'analyse des URL ou encore des historiques de visite de sites web. De manière générale, il apparaît que plus le nombre de critères est important, meilleure est la capacité du système à séparer les différentes classes.

Cependant, une préoccupation importante est de trouver le bon compromis entre les bons résultats (liés au nombre de critères) et le temps requis pour prendre une décision. Ainsi, de nombreux travaux tels que (Yang, Pedersen, 1997) recherchent l'automatisation de la sélection des critères les plus discriminants. *A fortiori*, le problème traité ici impose aux solutions proposées de pouvoir traiter en temps réel les nouveaux documents du flux. Ainsi le calcul d'un grand nombre de critères peut être un frein aux bonnes performances du système (en terme de temps et non de précision). De même des critères dont l'estimation des valeurs est trop coûteuse (en terme de temps toujours) sont à proscrire (on pensera par exemple à des analyses syntaxiques ou sémantiques poussées).

Nous proposons un ensemble de 40 critères communs à la résolution des deux sous-problèmes proposés ci-dessus. Ces critères se répartissent en trois catégories :

caractéristiques du document à proprement parler (contenu et structure), adéquation du contenu du document avec le profil de l'entité (son entrée dans la base de connaissances) et son contexte temporel. La section 2.1 présente ces critères et les intuitions qui y sont liées.

2.1. Critères pour la caractérisation d'un document en fonction de sa pertinence

2.1.1. Analyse du contenu des documents

La première source d'information est bien sûr le contenu du document lui-même. Nous proposons un ensemble de critères concernant le nombre de mentions de l'entité dans le document, leur position et la nature du document :

– nombre de mentions dans le document : un document dont le contenu est centré sur l'entité doit la mentionner de nombreuses fois. Le premier critère correspond donc à la probabilité d'observer l'entité en tirant un mot au hasard du document :

$$p(e|d) \propto tf(e,d) = \frac{c(e,d)}{\sum_{m \in d} c(m,d)} \quad (1)$$

où $c(e,d)$ est le nombre d'occurrences de e dans d, normalisé par rapport à la taille de d (la somme du nombre d'occurrences de chaque mot m) ;

– répartition des mentions : si le document est centré sur l'entité étudiée alors il est probable que les mentions de celle-ci soient réparties dans tout le texte. Pour mesurer cette répartition nous proposons d'utiliser comme critère la probabilité de tirer au hasard parmi toutes les phrases du document une phrase contenant l'entité :

$$p(s_e|d) \propto \frac{c(s_e,d)}{c(s,d)} \quad (2)$$

où $c(s_e,d)$ est le nombre de phrases s_e contenant e dans d et $c(s,d)$ le nombre de phrases contenues dans d ;

– présence de l'entité dans le titre : le précédent critère ne prenait pas en compte la position des occurrences de l'entité dans le document. Cependant de nombreux travaux, notamment en résumé automatique (Das, Martins, 2007), montrent que toutes les phrases n'ont pas la même importance. En particulier, (Edmundson, 1969) et (Kupiec *et al*., 1995) montrent que pour les documents journalistiques, le titre contient des informations essentielles sur les sujets abordés. Utiliser le contenu du titre s'est révélé pertinent dans de nombreuses applications comme la catégorisation de textes (Ko *et al*., 2002) ou la sélection des termes dans un processus de retour de pertinence simulé (pseudo-relevance feedback) (Lam-Adesina, Jones, 2001) et pour des types de documents divers. Ainsi, nous définissons un critère pour le titre, de manière similaire à celui pour le document, comme la probabilité d'observer l'entité dans le titre :

$$p(e|d_t) \propto tf(e,d_t) = \frac{c(e,d_t)}{\sum_{m \in d_t} c(m,d_t)} \quad (3)$$

où $c(e, d_t)$ est le nombre d'occurrences de e dans le titre du document d normalisé par rapport à la taille du titre d_t (la somme du nombre d'occurrences de chaque mot m) ;

– présence/absence d'un titre : (Shen *et al.*, 2004) utilisent les techniques de résumé automatique avec succès pour catégoriser des pages web. Des critères sont basés sur le titre. Cependant les pages web ne sont pas aussi bien formatées que les articles journalistiques ou scientifiques utilisés dans les références précédentes. D'après leur analyse près de 25 % des documents web ne contiennent pas de titre et il convient de séparer ces documents des autres (du moins de signifier cette différence). Nous proposons pour cela un critère booléen soulignant la présence ou non d'un titre pour un document donné;

– position des mentions : le titre n'est pas la seule partie du document à revêtir une importance particulière. (Baxendale, 1958) constate que pour 85 % des textes (sur un échantillon de 200 textes) les phrases contenant l'essentiel de l'information se retrouvent au début de ceux-ci et dans 7 % des cas à la fin. En résumé automatique une approche de référence largement utilisée consiste à former un résumé à partir des premières phrases du document (Edmundson, 1969) ou de chaque paragraphe (appelée *Lead method*) (Teufel, Moens, 1997). Dans cet esprit nous proposons un ensemble de critères correspondants au nombre de mentions de l'entité e par tranche de 10 % ou 20 % du texte (en nombre de mots) :

$$p(e|d, i) \propto tf(e, d_i) = \frac{c(e, d_i)}{\sum_{m \in d_i} c(m, d_i)} \tag{4}$$

où $tf(e, d_i)$ est le nombre d'occurrences de e dans une sous-partie du document d (déterminée par l'intervalle i) normalisé par rapport à sa taille (la somme du nombre d'occurrences de chaque mot m) ;

– homogénéité du document : est-ce que le document traité est centré sur un sujet ou en traite plusieurs ? Une estimation de cette homogénéité pourrait être faite à partir de l'analyse de l'importance des sujets latents présents dans le document (en ayant recours par exemple à l'allocation de Dirichlet latente (Blei *et al.*, 2003) ou l'analyse sémantique latente (Hofmann, 1999)). Cependant, ces traitements peuvent être coûteux en temps de calcul. À la place, nous proposons d'utiliser l'entropie de Shannon qui correspond à la quantité d'information délivrée par une source: plus elle émet d'informations différentes, plus l'entropie est grande. [6]. Elle est définie comme :

$$H_2(d) = -\sum_{m_i \in d} P_i \times log_2 P_i \quad ou\ P_i \propto \frac{c(m_i, d)}{\sum_{m_j \in d} c(m_j, d)} \tag{5}$$

6. voir http://www.cse.iitb.ac.in/čs626-460-2012/seminar_ppts/NLP_and_Entropy.pdf pour un aperçu de ses applications en traitement automatique des langues.

où $H_2(d)$ est l'entropie du document d, m_i un mot du document et P_i sa probabilité d'apparition dans le document ;

– taille du document : c'est un paramètre important dans nombre de mesures de similarité et approches de recherche d'information (voir par exemple (S. E. Robertson, Jones, 1997)) ainsi que dans de nombreuses approches de classification de document. Nous l'avons donc inclu à la liste des critères ;

– type du document : notre approche se veut générique et applicable à tout type de documents. Cependant, des différences importantes de style ou de structure existent. Pour notifier ces différences nous proposons l'ajout d'un critère booléen par type de document signifiant si le document est de ce type ou non (dans nos expériences nous aurons deux critères : *isSocial* pour les blogs ou forums et *isNews* pour les documents journalistiques).

2.1.2. Profil

Une vaste majorité des travaux sur l'association d'une mention à une entité (représentée par une entrée dans une base) recourent à des mesures de similarités entre le contexte de la mention et ce que nous pourrions appeler un profil de l'entité construit à partir des informations de la base de connaissances. Comme pour la tâche *TAC Knowledge Base Population (KBP)* nous nous plaçons dans le cas où la base de connaissances est Wikipédia. Cependant, les critères que nous proposons pourraient être facilement adaptés à d'autres bases de connaissances :

– similarité vectorielle : le critère le plus populaire dans les systèmes s'évaluant dans le cadre de KBP est l'utilisation d'une mesure de similarité entre un vecteur composé des mots du document mentionnant l'entité et un vecteur formé des mots contenus dans l'article Wikipédia supposé représenter l'entité (Ji *et al.*, 2011). Nous avons choisi la mesure du cosinus de l'angle entre les deux vecteurs car elle est symétrique et qu'une normalisation en fonction de la taille des documents est réalisable. Nous proposons tout d'abord deux critères : un premier correspondant à la valeur de cette similarité avec des vecteurs composés d'unigrammes et le second avec des vecteurs composés des bigrammes présents dans les documents. Il a été montré dans de nombreux travaux sur la mesure de la similarité sémantique entre deux mots que construire les vecteurs en n'utilisant que les termes présents autour des mots étudiés améliore les performances. Dans cette optique nous ajoutons un troisième critère pour lequel le vecteur pour le document ne sera pas construit sur l'intégralité de son contenu mais sur la concaténation de fenêtres de 50 mots autour de chaque mention de l'entité dans le document ;

– mesure de la présence d'entités liées : la présence dans un document d'une entité que l'on sait liée à l'entité étudiée permet à la fois de désambiguïser la mention (peu de chance que deux homonymes soient en relation avec la même entité) mais aussi

peut être révélatrice de l'importance de l'information contenue dans le document. De nombreuses approches reposent sur leur utilisation, en particulier (Liu, Fang, 2012) et (Araujo *et al.*, 2012) à TREC KBA 2012.

Les entités liées sont retrouvées de deux manières :

- toutes les entités trouvées dans la page Wikipédia à l'aide d'un outil de reconnaissance des entités nommées (dans notre cas Stanford NER [7]);

- les entités pointées par la page (peut-être plus importantes que les autres).

À chaque entité re_i trouvée, un poids est associé en fonction du nombre d'occurrences :

$$w(re_i, e) = \frac{tf(re_i, e)}{\sum_{re \in RE(e)} tf(re, e)} \tag{6}$$

où RE est l'ensemble des entités liées à l'entité e.

À un document donné un score est attribué en fonction du nombre de mentions d'une entité liée et son poids :

$$TF(re, e, d) = \sum_{re_i \in RE(e)} c(re_i, d) \times w(re_i, e) \tag{7}$$

où $c(re_i, d)$ correspond au nombre de mentions de re_i dans d et $w(re_i, e)$ à l'importance de l'entité re_i pour e;

– taille de l'entrée dans la base de connaissances : la taille en nombre de mots de la page Wikipédia. À l'instar de la taille du document étudié ce critère pourrait jouer un rôle. Par exemple, un document Wikipédia long va traiter de nombreux sujets différents et ainsi les mesures de similarité pourraient être biaisées.

2.1.3. Analyse temporelle

De nombreux facteurs comme la périodicité des centres d'intérêts des utilisateurs (Yom-Tov, Diaz, 2011), la répétition d'une même information dans de nombreux documents, la complémentarité de différentes sources d'information comme les sites de nouvelles journalistiques et les blogs (König *et al.*, 2009), les réseaux sociaux (Dong *et al.*, 2010) ou encore la nouveauté rentrent en jeu (Del Corso *et al.*, 2005). (Radinsky *et al.*, 2013) présentent une vaste vue d'ensemble des résultats et applications de l'analyse et de la prise en compte de la composante temporelle pour la recherche d'information.

La notion de temps permet une mise en contexte d'un document en permettant par exemple des comparaisons avec des documents de la même période ou de périodes antérieures. Nous souhaitons estimer si la période de parution d'un document est propice

7. http://nlp.stanford.edu/ner/index.shtml

à l'apparition de documents pertinents pour une entité donnée. Pour cela nous proposons plusieurs critères pour détecter des tendances ou des sursauts (ou pics, *burst* en anglais) dans le volume de documents citant l'entité.

La détection de pics de popularité d'un sujet donné est un problème ouvert. L'une des difficultés est la variété des phénomènes pouvant être observés et les différents schémas que peuvent suivre les changements au cours du temps. De plus les phénomènes ne sont pas tous observables à la même échelle de temps : (Rattenbury *et al*., 2007) montrent qu'autant une échelle d'une heure ou d'une journée permettent de détecter un évènement, autant certains phénomènes n'émergent clairement qu'au bout d'une semaine ou plus. Nous proposons ainsi des critères pour chacun de ces intervalles, souhaitant capturer un maximum d'information :

– pondération du nombre de mentions : le premier critère considéré est une pondération du nombre d'occurrences dans un document en fonction du nombre de documents mentionnant l'entité dans l'heure précédente. C'est donc une sorte de TF.IDF qui est mesurée mais l'IDF n'est calculé que pour un sous-ensemble des documents trouvés jusqu'alors : ceux des 60 dernières minutes ;

– nombre de documents sur 24 heures : le second critère est le nombre de documents contenant une occurrence de l'entité dans les précédentes 24 heures. Ce critère équivaut à l'IDF de l'entité dans une collection constituée de tous les documents des 24 dernières heures. Nous souhaitons ainsi mesurer l'apparition d'une nouvelle tendance ;

– nombre de documents sur 7 jours : il est difficile d'estimer une évolution sans prendre en compte une échelle de temps plus large que 24 heures. Cette mesure est donc aussi effectuée sur 7 jours ;

– variance et écart type du nombre de documents par jour : afin de refléter le caractère commun ou exceptionnel du nombre de documents, nous mesurons le nombre moyen de documents par jour les 7 derniers jours ainsi que sa variance. (Diaz, 2009) utilise un critère similaire pour exacerber les différences entre chaque période ;

– nombre de titres avec une mention sur 7 jours: le titre est *a priori* un indicateur fort de la thématique d'un document. Ainsi, il semble naturel de mesurer l'évolution de la présence de l'entité dans les titres des documents des 7 jours précédant le document analysé.

2.2. Détection des documents centrés sur une entité dans un flux

Pour déterminer si un nouveau document du flux documentaire est pertinent au regard d'une entité donnée, deux classifieurs, utilisés séquentiellement, sont construits et exploitent les critères précédents. Le premier sélectionne les documents pertinents, et le second distingue, parmi les documents jugés pertinents, ceux qui sont vitaux. Dans les expérimentations reportées ici, nous avons opté pour l'utilisation de forêts d'arbres décisionnels. Dans nos diverses expériences ils obtiennent des résultats com-

pétitifs avec d'autres classifieurs (comme des SVMs) et ils permettent en outre une analyse aisée de la hiérarchie des critères utilisés dans la prise de décision.

3. Cadre expérimental et résultats

Dans cette section, nous commençons par présenter le cadre d'évaluation proposé par la tâche KBA à TREC 2012 et les adaptations de notre système, puis, dans un second temps, les résultats officiels.

3.1. TREC KBA 2012

Dans le cadre de cette campagne d'évaluation, un corpus a été élaboré. Il est composé de trois catégories de documents pour un total de près de 9 To de données, soit environ 500 millions de documents. Les trois catégories de documents sont :

- *social* : ensemble de documents provenant de blogs et forums ;
- *web* : documents web provenant de la base de Bitly [8] ;
- *presse* : documents journalistiques.

Ces documents ont été collectés entre octobre 2011 et avril 2012. A chaque document sont associées la date et l'heure précise de sa publication. Le corpus est divisé en deux parties : d'octobre à décembre 2011 pour l'entraînement, et de janvier à avril 2012 pour le test.

Tableau 1. Corpus KBA 2012 : nombre de documents et taille par catégorie

	Presse	Web	Social
# docs	134 625 663	5 400 200	322 650 609
taille	8072Go	350Go	531Go

Pour le test, 29 entités ont été sélectionnées pour leur difficulté et leur degré d'ambiguïté (voir la liste figure 2) ainsi que pour leur faible nombre d'occurrences dans le corpus (l'objectif étant de ne pas chercher à cibler des entités trop populaires ni, à l'inverse, des personnes pour lesquelles trop peu d'information sont disponibles sur le Web). Les annotateurs humains devaient associer l'une des trois classes suivantes aux documents : non pertinent, intéressant, vital. La mesure retenue est la F-mesure (moyenne harmonique entre précision et rappel).

3.2. Associer un score aux documents

L'évaluation dans le cadre de KBA impose aux systèmes de fournir en sortie, pour chaque entité, une liste de documents ordonnés. Notre approche, telle que présentée en section 2, associe une étiquette à chaque document parmi "non pertinent", "pertinent"

8. https://bitly.com/

et "vital". Cependant, les classifieurs utilisés associent à chaque prise de décision un score entre 0 et 1 faisant office de score de confiance. Utilisant cette information, nous procédons de la sorte pour associer un score à un document :

$$S_1(d_i) = \begin{cases} s(d_i, c_{np,pv}) \times s(d_p, c_{p,v}) \text{ si } s(d_p, c_{np,pv}) >= 0,5 \\ \text{retiré de la liste sinon} \end{cases} \quad (8)$$

où le score du document d_p est donné comme le produit du score renvoyé par le classifieur $c_{np,pv}$ (départageant les non pertinents (np) des pertinents (p) et vitaux (p)) et par le classifieur $c_{p,v}$ [9].

3.3. Résultats

Dans cette sous-section, sont présentés les résultats dans le cadre de la tâche KBA à TREC 2012. L'évaluation porte sur la capacité des systèmes à trouver dans le flux les documents dits "vitaux".

3.3.1. Analyse globable

La figure 1 présente les résultats de toutes les soumissions des différents participants à la tâche KBA 2012 ainsi que le score obtenu par notre proposition pour retrouver les documents vitaux. En noir notre approche et en gris les approches de référence : *ref-svm* utilisant un classifieur par entité avec pour critères les mots présents dans les documents, *ref-el* mesurant la présence d'entités liées à l'entité étudiée (extraites de Wikipédia) dans le document et *ref-match* pour laquelle tous les documents contenant le nom de l'entité ou une de ses variantes (surnom, abréviation etc.) sont retournés. On pourra tout d'abord remarquer grâce à cette figure que la tâche est difficile : la plupart des systèmes de filtrage ont des performances inférieures au système de référence consistant à ne filtrer les documents qu'en fonction de la présence/absence de l'entité (points sous la ligne horizontale). Notre approche, point en haut à droite sur la figure, obtient des résultats supérieurs à ceux des autres participants.

Le tableau 2 compare les résultats de notre approche à ceux des systèmes de référence mais aussi de la médiane et la moyenne de tous les participants à la tâche. La construction des forêts d'arbres de décision est un processus induisant une variabilité non négligeable dans la qualité de la classification. Afin de tenir compte de cette variabilité, nous présentons la moyenne de 50 évaluations différentes. Notre approche avec un score moyen de 0,382 obtient un résultat légèrement supérieur à celui du meilleur système à KBA 2012 ($0,359$, $+6$ %) et largement supérieur à la médiane ($0,289$, $+32$ %) et la moyenne ($0,220$, $+73$ %).

9. Différentes formules, prenant en compte ou non les résultats des deux classifieurs ont été envisagé, elles obtiennent des résultats équivalents. Dans un souci de clarté, nous ne reportons ici que les résultats obtenus pour cette fonction de score, plus intuitive et simple.

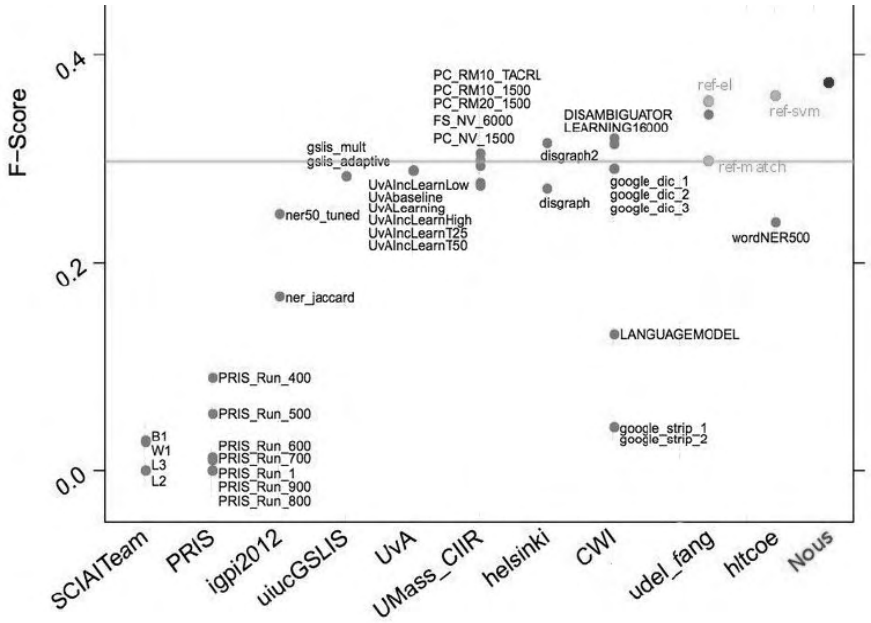

Figure 1. Résultats officiels pour les documents vitaux à KBA 2012 pour la F-mesure. Chaque point sur une verticale correspond à différents paramétrages d'un même système

Tableau 2. F1-mesure pour la détection de documents vitaux de notre approche (score moyen de 50 évaluations) contre les systèmes de référence, la médiane et la moyenne à KBA 2012

Approche	F1-mesure	Approche	F1-mesure
Nous	0,382	ref-match	0,297
ref-svm	0,359	Médiane	0,289
ref-el	0,355	Moyenne	0,220

3.3.2. Impact de la taille du corpus d'apprentissage

La méthode *ref-svm* a obtenu les meilleurs résultats pour la tâche KBA 2012 mais nécessite un corpus d'apprentissage par entité étudiée. Notre approche, obtenant des résultats supérieurs, n'a recours qu'à un seul jeu de deux classifieurs entraînés une fois pour l'ensemble des entités. Nous prétendons que cette approche capture des caractéristiques permettant de déterminer la pertinence d'un document pertinent, indépendamment de l'entité étudiée. Nous prétendons de surcroît que si aucun corpus d'apprentissage n'est associé à une nouvelle entité à traiter les performances n'en sont que faiblement impactées. Enfin nous pensons qu'un faible nombre d'exemples est nécessaire pour entraîner le système.

Pour démontrer le bien-fondé de ces affirmations, nous avons évalué les performances de l'approche pour la tâche KBA 2012 en utilisant une méthode proche d'une validation croisée. Nous divisons l'ensemble des entités en k échantillons de manière à ce que chaque échantillon contienne une et une seule entité. La tâche est ensuite effectuée pour chaque échantillon en entraînant les classifieurs sur les exemples correspondant aux (k-1) autres échantillons. Le résultat présenté dans le tableau 3 sous le nom *loocv (leave-one-out cross validation)* montre que sous cette configuration notre proposition obtient un score de F1-mesure moyen de 0,361. Ainsi, nous pouvons affirmer que même privé de corpus d'apprentissage spécifique à l'entité traitée notre système est toujours performant (avec un résultat légèrement supérieur à celui du meilleur système officiel de KBA 2012). Ce bon résultat semble aussi appuyer notre conclusion quant à la capacité de la méthode proposée à capturer de manière générale les caractéristiques d'un document contenant des informations vitales sur une entité.

Tableau 3. F1-mesure en validation croisée pour différents nombre d'échantillons et pour le filtrage des documents vitaux. Les résultats présentés sous le nom de crossK sont la moyenne des résultats obtenus pour cinquante tirages aléatoires des éléments présents dans chacun des k échantillons.

Run	F-mesure	Run	F-mesure
Tout	0,382	cross5	0,350
LOOCV	0,361	cross3	0,354
ref-svm	0,359	cross2	0,339
cross10	0,355	ref-match	0,297

Pour évaluer la robustesse de notre approche, différentes configurations de type validation croisée ont été réalisées. Précédemment, autant d'échantillons que d'entités étaient utilisés. Nous allons réduire le nombre d'échantillons afin de diminuer le nombre d'exemples utilisés pour évaluer un échantillon donné. Les différentes valeurs de k envisagées sont $k \in \{10, 5, 3, 2\}$. Pour chaque valeur de k, cinquante évaluations ont été faites. Pour chacune de ces évaluations, l'assignation des entités aux échantillons a été faite de manière aléatoire. Les valeurs reportées dans le tableau 3 sous les noms de *crossK* sont la moyenne des résultats obtenus pour chacune des cinquante évaluations pour un k donné.

Naturellement, les résultats décroissent à mesure que le nombre d'exemples d'entraînement diminue. Une exception cependant pour $k = 3$ qui est très légèrement supérieur au cas $k = 5$. Une explication peut être que le nombre d'exemples associés à chaque entité n'est pas homogène. Ainsi certaines entités ont beaucoup d'exemples qui leur correspondent et leur exclusion pénalise grandement les systèmes.

Les performances de notre système restent très largement supérieures à la plupart des systèmes ayant participé à KBA 2012. Plus précisément, pour $k = 10$, notre système se placerait en seconde position de la compétition, tandis que pour $k \in \{5, 3, 2\}$ les résultats permettraient à l'approche de se classer en troisième position, loin devant la médiane ou encore l'absence de filtrage (méthode *ref-match*). Ceci montre la capa-

cité de l'approche proposée à généraliser rapidement, à l'aide des critères proposés, ce qui distingue un document important d'un autre.

3.3.3. Analyse à l'échelle des entités

Nous avons vu que notre approche obtient en moyenne de bons résultats sur l'ensemble des 29 entités. Pour préciser les difficultés que notre approche rencontre ainsi que les succès de celle-ci, nous proposons désormais d'étudier les résultats au niveau de chaque entité.

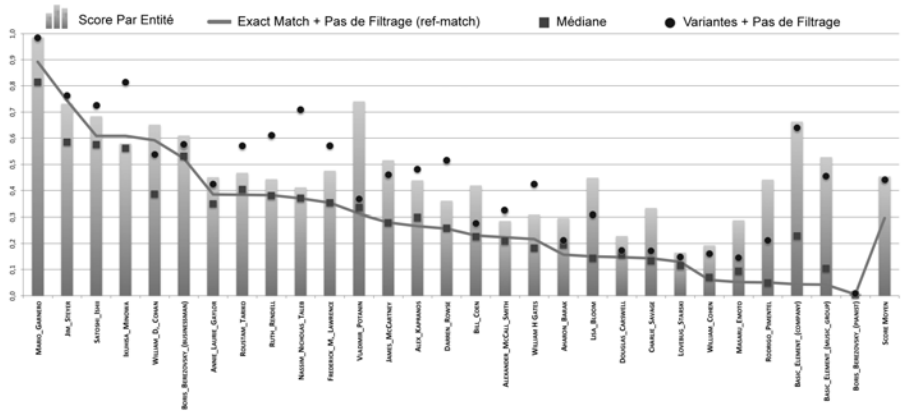

Figure 2. F1-mesure pour le filtrage des documents vitaux entité par entité. Sont représentés : les résultats de notre approche (barres), le score médian (les carrés), score obtenu en récupérant tous les documents contentant le nom de l'entité (la ligne) et des variantes du nom (les ronds)

La figure 2 présente le résultat de notre approche entité par entité ainsi que la médiane des approches concurrentes à la tâche KBA 2012 (carrés). L'approche de référence *ref-match* (*i.e.* récupération des documents contenant exactement l'entité et sans filtrage) est représentée par la ligne. Pour s'assurer un large rappel, les annotateurs de la tâche ont tout d'abord recherché des variantes d'écriture du nom de l'entité. Tous les documents contenant ces variantes ont été évalués. Les ronds représentent les scores qui seraient obtenus par un système capable de retrouver tous ces documents et n'appliquant aucun filtrage. Les barres correspondent au score obtenu pour une entité. Les scores pour chaque entité correspondent au score moyen obtenu pour cinquante évaluations.

On peut tout d'abord constater que le score médian des différents participants à la campagne d'évaluation pour chaque entité est généralement très proche de celui du système *ref-match*. Ainsi, la plupart des systèmes effectuent en moyenne un filtrage des mauvais documents (ils obtiennent des moins bons résultats qu'une méthode qui consiste uniquement à extraire, sans les différencier, les documents dans lesquels le nom de l'entité apparaît).

Pour toutes les entités, le filtrage et le classement effectués par notre approche améliorent le résultat en termes de F-mesure par rapport à ce système de référence ainsi que par rapport à la médiane. Pour certaines entités (comme par exemple *Vladimir Potanin*, *Lisa Bloom* et *Basic Element*) une importante différence en notre faveur existe. Pour les entités les plus faciles (du point de vue ratio vitaux/autres élevé et score médian), notre approche de filtrage ne permet guère d'améliorer les résultats. En revanche, pour les entités plus difficiles, notre approche de filtrage permet d'éliminer des documents non pertinents.

La figure 3 présente le nombre de documents jugés vitaux par entité et par heure dans le corpus. On peut remarquer que pour certaines entités telles que *James McCartney*, *Mario Garnero* ou encore *Boris Berezovksy (homme d'affaire)* la présence de pics dans le nombre de documents semblent coïncider avec de bons résultats. Cependant l'existence de tels pics ne garantit pas de bonnes performances comme on peut le voir pour *William H Gates* et à l'inverse de bons résultats ne semblent pas forcément liés à ces pics (voir *Vladimir Potanin*) et c'est la force de notre approche qui permet de capturer différents types de comportements.

Les résultats obtenus pour *Basic Element (le groupe musical)* et *Boris Berezovksy (le pianiste)* peuvent en revanche s'expliquer principalement par l'existence d'homonymes qui sont de plus très présents dans le corpus. La présence de noms communs dans certains noms comme pour *Basic Element* ou *Lovebug Starski* peuvent provoquer une baisse des résultats à cause du grand nombre de documents retrouvés (raison pour laquelle ref-match obtient de faibles scores pour ces entités).

Cependant, il est difficile de trouver un schéma commun entre plusieurs entités car leur nombre est trop faible alors que leurs profils sont très variés. On peut trouver dans (Frank *et al*., 2012) le nombre de documents annotés pour chaque entité et les proportions de documents vitaux, pertinents, neutres et déchets. On constate alors qu'il n'existe aucune paire d'entités présentant des valeurs similaires pour ces différents paramètres. Ceci, ajouté à la variété des types d'entités proposées (homme d'affaire, musicien, entreprise *etc.*) rend les analyses par entité ardues.

Enfin, l'utilisation de variantes du nom de l'entité pour trouver des documents candidats pourrait améliorer les performances. Sur la figure 2, les points représentent le score qui serait obtenu en considérant tous les documents contenant le nom de l'entité testée ou un alias comme vitaux (*i.e.* sans filtrage). En comparant ces résultats avec ceux du système *ref-match* (représenté par la courbe) pour lequel seuls les documents contenant le nom exact de l'entité sont considérés, on constate une amélioration systématique et souvent importante du F1-score. Ainsi, il apparaît que les documents contenant une variante du nom de l'entité sont en moyenne plus pertinents que ceux contenant le nom de l'entité uniquement (car en proportions moyennes moins nombreux et pourtant impactant fortement à la hausse les résultats). Cette différence de pertinence moyenne s'explique peut-être par le fait que les variantes sont plus spécifiques à l'entité comme par exemple un surnom. De même, en moyenne, les résultats sont équivalents à ce que peut faire de mieux notre approche.

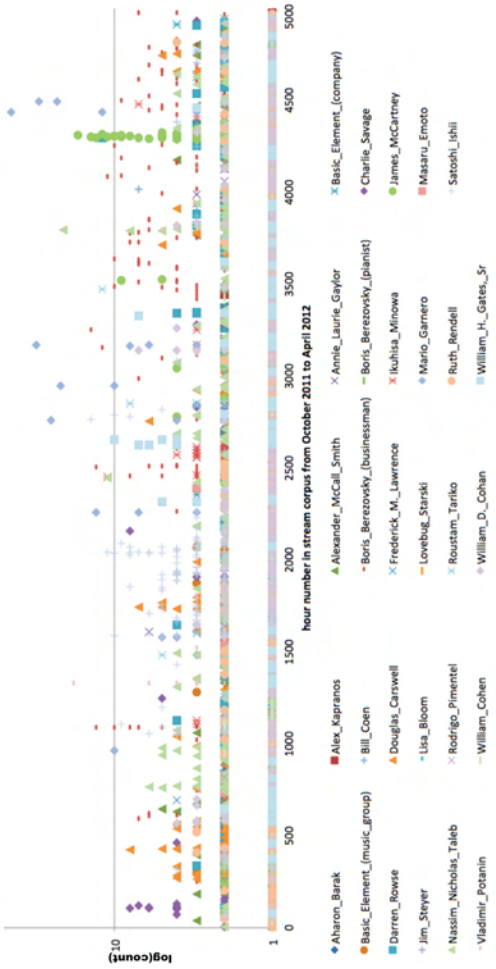

Figure 3. Nombre de documents par entité et par heure pour l'intégralité du corpus

3.3.4. Impact des critères

L'utilisation de forêts d'arbres de décision (*Random Forest* (Breiman, 2001)) comme approche englobante (Kohavi, John, 1997) pour sélectionner et mesurer l'importance des critères s'est montrée efficace dans de nombreux travaux comme la sélection de gènes (Díaz-Uriarte1, Andrés, 2006) ou encore les relations entre molécules pharmaceutiques (Svetnik *et al.*, 2004).

Lors de la construction d'un arbre de la forêt, des exemples sont exclus aléatoirement. Ceci permet de limiter le sur-apprentissage et d'estimer les performances de chaque arbre en les évaluant sur les exemples exclus lors de leur construction. Ce procédé permet aussi d'estimer l'importance de chaque critère dans les prises de décision :

– pour un arbre donné, les exemples exclus sont traités par celui-ci dans une phase de test. Le nombre d'exemples correctement associés à leur classe est compté ;

– toujours pour ce même arbre et pour un critère donné, les valeurs de chacun de ces exemples sont permutées et les exemples sont à nouveau classés par l'arbre. Le nombre d'associations correctes est là encore relevé ;

– pour un arbre et un critère donnés, la différence entre les deux quantités précédentes est un estimateur de l'importance du critère (Strobl *et al.*, 2008) ;

– l'importance globale d'un critère correspond à la moyenne de son importance sur chaque arbre. Ainsi, un critère est considéré comme d'autant plus important que la permutation des valeurs de ce critère entre les exemples engendre une diminution importante des performances de chaque arbre.

Nous présentons, pour chacun des critères, l'importance moyennée sur 50 forêts d'arbres de décisions. L'implémentation utilisée est celle du paquetage *party* (*A Laboratory for Recursive Partytioning*) de l'environnement logiciel *R* [10] avec cinq critères candidats pour la séparation des données à chaque nœud et une profondeur maximale de 50 pour chaque arbre.

La figure 4 présente l'importance moyenne de chaque critère pour séparer les documents pertinents des vitaux (p/v) pour le corpus d'entrainement (en gris foncé) et celui de test (en gris clair).

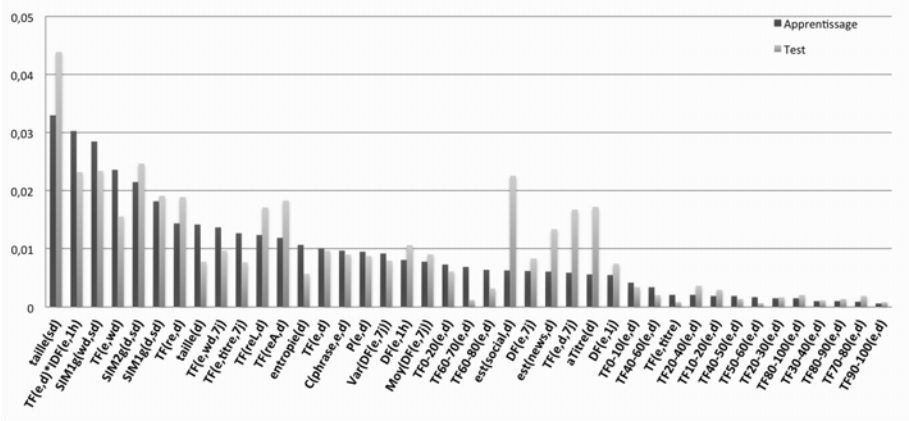

Figure 4. Importance des critères moyennée sur 50 forêts d'arbres de décisions appris. En gris clair pour le corpus d'apprentissage de KBA 2012 et en gris foncé pour celui de test

Sur cette figure, nous pouvons remarquer que les cinq critères les plus discriminants correspondent à la similarité du document avec le document source (la page Wikipedia de l'entité) mais aussi à la présence des entités liées dans le document. On

10. http://www.r-project.org

remarquera que la présence d'entités liées permet efficacement de déterminer quels sont les documents mentionnant l'entité et qui sont donc pertinents mais permettent moins de séparer les documents pertinents et les documents vitaux. Enfin, le nombre d'occurrences de l'entité dans le document s'avère un critère important pour différencier les documents pertinents entre eux. Au-delà du rang 6, les critères semblent peu utiles pour séparer efficacement les documents. Pour confirmer cette observation nous avons réévalué le système en utilisant uniquement ces critères. La F1-mesure obtenue est de $0,374$ contre $0,382$ avec la totalité des critères et pour rappel $0,289$ de score moyen de l'ensemble des participants.

Cette figure nous réserve aussi quelques surprises, en particulier l'impact presque nul de la prise en compte du titre. La présence ou l'absence d'un titre pour un document ne semble pas discriminante et la présence de l'entité dans le titre est le critère qui l'est le moins alors même qu'il ne semble être fortement corrélé avec aucun autre critère. Une analyse des documents du corpus d'apprentissage montre en effet que seulement 53 % des documents contenant l'entité dans le titre s'avèrent pertinents ou vitaux.

La position des occurrences dans le document semble ne pas avoir de lien non plus avec la classe d'un document alors qu'en résumé automatique une approche basique mais performante consiste à considérer les phrases en début et fin de document comme particulièrement révélatrice du contenu. De plus, les différentes parties semblent toutes être autant discriminantes (ou peu discriminantes) les unes que les autres. Une exception existe : la prise en compte des occurrences dans les 20 premiers pourcents du document qui est légèrement plus décisive. Cependant la manière dont nous considérons la position des occurrences dans le texte est à mettre en cause dans ces résultats. (Berendsen *et al*., 2012) le démontrent en utilisant les positions pour calculer le nombre de mots séparant la première occurrence de l'entité dans le document avec sa dernière occurrence. Ce critère s'est révélé extrêmement déterminant. Leur résultat est cependant à prendre avec précaution car il n'est pas mentionné que l'importance de ce critère est moyennée sur plusieurs expérimentations et ce bon résultat pourrait être le fruit du hasard. (Balog *et al*., 2013) indique toutefois un résultat similaire, mais là aussi apparemment issu d'une seule expérimentation.

Le fait que ces résultats sur l'importance relative des critères ne semblent pas conformes à ceux bien connus en résumé automatique provient peut être de la nature des documents plus que d'une différence entre les tâches. La plupart des travaux en résumé concernent des documents journalistiques dont le style et la structure diffèrent des entrées de blogs ou de forums qui constituent les deux tiers de la collection KBA.

Pour le corpus de test la première constatation est que l'ordre d'importance des critères est peu altéré par rapport à celui obtenu sur le corpus d'apprentissage : les 10 critères les plus pertinents sont les mêmes à deux exceptions près et il en va de même pour les critères les moins pertinents (position dans les documents et nombre d'occurrences dans le titre).

Parmi les exceptions, on notera tout particulièrement le cas du critère *estSocial(d)* qui s'avère être le 4^e critère le plus influent pour la caractérisation des documents vitaux (contre 14^e sur le corpus d'apprentissage). Nous expliquons cette hausse par la présence d'un troisième type de document (pages web autres que blogs et forums) qui est absent du corpus d'apprentissage et donc par la nécessité de les différencier. Cette explication est appuyée également par une hausse, bien que légère, pour le critère *estNews(d)*. Enfin un troisième critère allant dans cette direction est *aTitre(d)* permettant là aussi probablement une meilleure séparation des types de documents pouvant présenter des caractéristiques différentes.

Le nombre de documents mentionnant l'entité en une heure est un critère nettement plus important pour séparer les non pertinents des autres sur le test que sur l'apprentissage. Il semblerait donc que des phénomènes non (ou moins) présents dans le corpus d'apprentissage apparaissent dans le test.

La dernière question en suspens est : y a-t-il des critères plus importants que d'autres en fonction du type des document étudiés ? L'importance des critères pour les documents sociaux uniquement est similaire à celle obtenue sur l'ensemble du corpus (hormis bien sûr en ce qui concerne *estSocial(d)* et *estNews(d)*). La prépondérance des documents sociaux dans le corpus KBA 2012 explique peut-être cette similitude. Pour les documents journalistiques, on peut noter des différences plus importantes. Deux critères voient leur importance réduite à néant : la présence ou l'absence d'un titre ainsi que le nombre de mentions de l'entité dans le titre. Cela confirme que ces critères sont principalement utiles pour distinguer les types de documents et non pour distinguer le degré de pertinence. D'autres critères, à l'inverse, se révèlent très discriminants : la présence dans les 10 ou 20 premiers pourcents du document. Ce résultat est cette fois en accord avec les travaux en résumé automatique et confirme que ces critères ne sont pas transposables à d'autres types de documents.

4. Expérimentations sur KBA 2013

4.1. De KBA 2012 à KBA 2013

En 2013, la tâche KBA a été reconduite et redéfinie sur certains points: le corpus, les entités et les classes de documents. En ce qui concerne le corpus, en 2013 celui-ci était composé de plus d'un milliard de documents (11 948 heures de flux) contre 400 millions en 2012 (4 973 heures). Le nombre d'entités sélectionnées a augmenté : 141 entités en 2013 contre 29 en 2012.

Par ailleurs, en 2013, les entités sélectionnées n'étaient pas forcément associées à une page Wikipedia et certaines sont des identifiants de la plateforme Twitter que l'on peut toujours associer au nom affiché sur Twitter (ex. `@urben00` alias Brent Faulkner). Ce changement est très important car le système mis en place pour KBA 2012 utilise la page Wikipedia associée à l'entité pour calculer certaines des caractéristiques utilisées pour la classification. Le nombre de documents annotés est en revanche beaucoup moins grand qu'en 2012 par rapport au nombre d'entités (voir table

4). On remarque qu'en 2013, il y a beaucoup moins de documents annotés, même pour les classes les plus importantes (*Relevant/Useful* et *Central/Vital*). On remarque également que le nombre de documents moyen par entité et par classe est beaucoup moins élevé, rendant la tâche d'autant plus difficile.

Tableau 4. Répartition des documents annotés pour l'entraînement par classe de document (1) et par entité (2) pour KBA2012 et KBA2013

Classe	**#Docs**[1]		**#Docs/Entity**[2]	
	2012	2013	2012	2013
Garbage	8467	2176	284	20
Neutral	1584	1152	73	11
Relevant/Useful	5186	2293	181	20
Central/Vital	2671	1718	92	19
Total	17482	7222		

Le dernier changement important en 2013 concerne les classes de documents. Les deux classes *relevant* et *central* ont été respectivement remplacées par *useful* et *vital*. Ce n'est pas simplement le nom qui change mais bien la définition des classes. Un document *useful* est un document qui parle essentiellement de l'entité cible mais qui relate des faits utiles dans la construction d'une biographie par exemple. Les documents vitaux quant à eux concernent essentiellement l'entité cible et, en plus, contiennent une information nouvelle sur cette entité.

4.2. Résultat de l'expérimentation

En 2013, nous avons testé quatre variantes de notre système différents parmi lesquelles une variante de celui de 2012 :

1. *"Single"* entraîne un seul classifieur sur les quatre classes de documents (Garbage, Neutral, Useful, Vital) ;

2. *"KBA 2012"* est composé de deux classifieurs utilisés en cascade. Chacun s'entraîne sur 2 classes. Le premier donne une classe parmi *Garbage/Neutral* et *Useful/Vital*. Le second classifieur donne une classe parmi *Useful* et *Vital* ;

3. *"VitalvsOthers"* entraine un classifieur qui apprend à distinguer seulement la classe *Vital* contre toutes les autres classes (*others*);

4. *"Combine"* exploite les scores donnés par les trois classifieurs précédents afin de trouver la meilleure combinaison possible.

Dans l'évaluation de KBA 2013, il y a deux aspects pris en compte : la recherche d'information (filtrage initial) et la classe attribuée à un document trouvé, donnant ainsi un résultat global pour la réalisation de la tâche complète. Cependant, pour une analyse plus fine nous évaluons, à l'aide de matrices de confusion, les scores obtenus par nos différentes variantes en ne prenant en compte que la classification. Les documents qui n'ont pas été classés, car ils n'ont pas été trouvés par notre système, ne seront pas comptabilisés comme pénalisant. Nous avons ensuite calculé les matrices

de confusion pour tous les scores de confiance $s \in \mathbb{N}]0 - 1000]$ pour enfin garder le meilleur compromis entre précision et rappel. Les résultats inscrits dans le tableau 5, montrent que les différentes variantes présentent des résultats complémentaires : une plus grande précision pour *VitalVsOther* au détriment du rappel qui est plus faible que pour toutes les autres variantes. La version *KBA 2012* quant à elle offre le rappel le plus élevé mais également la précision la plus faible. Enfin, la variante Single, qui utilise les quatre classes pour effectuer la classification, obtient la meilleure F-mesure. Le système combinant les scores de tous les classifieurs (*combine*) permet quant à lui d'arriver à un compromis entre précision et rappel. Il n'est pas étonnant que ce système n'obtienne pas le meilleur score, sachant qu'il essaye de tirer profit de tous les autres. Cependant, il est assez proche en terme de F-mesure du meilleur système et il offre par ailleurs une meilleure précision avec une perte en rappel qui est moindre.

Tableau 5. Résultats de la classification sans prendre en compte l'étape initiale de filtrage

Système	**Precision**	**Rappel**	**F-Mesure**
Single	.569	.406	**.474**
KBA2012	.475	**.436**	.455
VitalVsOthers	**.725**	.323	.447
Combine	.619	.368	.461

La figure 5, générée à partir des sorties de l'outil d'évaluation officiel, montre que la combinaison des scores est plutôt payante: la précision augmente lorsque le seuil de confiance est élevé (le plus à gauche), ce qui n'est pas le cas pour les autres variantes de notre système.

Pour terminer, nous avons calculé l'importance des critères pour la variante de notre système qui obtient la meilleure précision (*VitalVsOthers*) afin d'estimer l'impact des changements entre les évaluations 2012 et 2013. La figure 6 montre une évolution majeure concernant le critère temporel qui compte le nombre de documents comportant une mention de l'entité durant les dernières 24 heures. Cela permet de conforter l'idée que la temporalité a toute sa légitimité dans ce type de tâche et que sa prise en compte constitue une perspective majeure de nos travaux.

5. Conclusions et perspectives

Nous avons présenté une méthode supervisée pour estimer la degré d'intérêt d'un nouveau document dans un flux au regard d'une entité présente dans une base de connaissances. Cette approche est basée sur des critères tels que le nombre de mentions dans le document, la présence ou non d'entités liées ou encore différents indices mesurant l'ampleur de l'actualité de l'entité dans la période d'apparition du nouveau document. La pertinence de notre approche a été validée par de bons résultats obtenus dans le cadre d'une participation à la tâche KBA de la campagne d'évaluation TREC 2012. Elle nous a permis de mesurer la justesse de nos intuitions dans le choix des critères utilisés même s'il semble que des approches plus simples puissent s'avérer plus

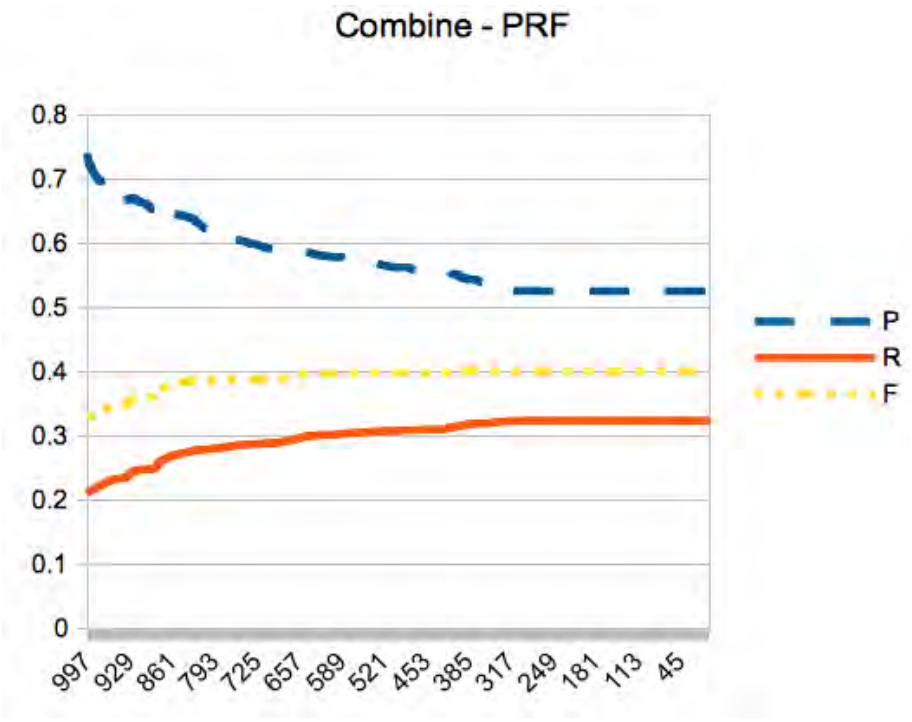

Figure 5. Variation de la précision, du rappel et de la F-mesure en fonction du score de confiance de l'approche Combine

performantes. Cela souligne l'intérêt et la difficulté de la tâche mais aussi la marge de progression, importante pour le futur.

L'un des principaux résultats et que certains critères, largement répandus en recherche d'information et résumé automatique, se sont révélés beaucoup moins discriminants qu'escompté (la position des occurrences des entités dans les documents, la présence ou non de l'entité dans le titre...) même si leur pouvoir discriminant pourrait être amélioré en normalisant les valeurs entre entités. Nous pensons également que les critères liés aux temps peuvent certainement avoir un rôle important à jouer dans la classification. Cependant, leur utilisation n'est pas encore maîtrisée et il reste beaucoup de questions sur lesquelles nous aimerions nous investir : les critères temporels doivent-ils être utilisés pour toutes les entités ? Certaines entités en tireraient plus de profit que d'autres ? Comment le déterminer ?

Certaines entités sont susceptibles d'évoluer au cours du temps. Nous pensons travailler sur la manière de mettre à jour l'entrée de l'entité dans la base au cours du temps. Cela permettrait entre autres d'opérer des connexions avec les problématiques liées à la population automatique de bases de connaissances, notamment dans le cadre de la campagne d'évaluation "Knowledge Base Population" à TAC.

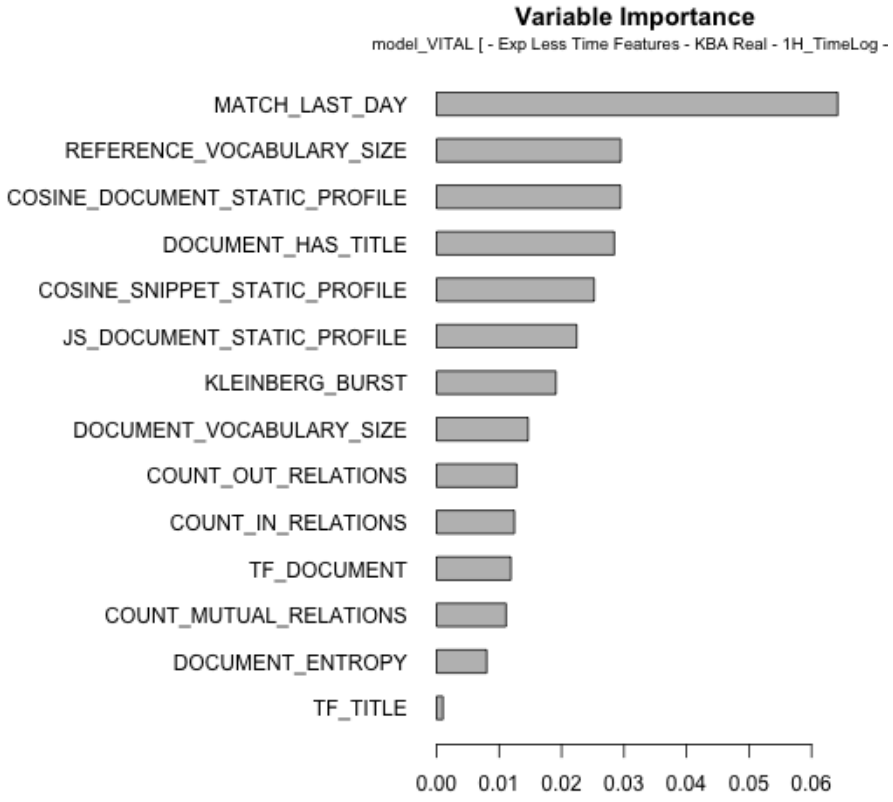

Figure 6. Classement des critères selon leur importance dans la classification pour la stratégie VitalvsOthers sur les données KBA 2013

Bibliographie

Araujo S., Gebremeskel G., He J., Bosscarino C., Vries A. de. (2012). Cwi at trec 2012, kba track and session trac. *Proceedings of The 21th Text Retrieval Conference (TREC)*.

Artiles J., Li Q., Cassidy T., Tamang S., Ji H. (2011). Cuny blender tac-kbp2011 temporal slot filling system description. *Proceedings of the Fourth TAC*.

Asahara M., Matsumoto Y. (2003). Japanese named entity extraction with redundant morphological analysis. *Proc. of the Human Language Technology conference - ACL*.

Atkinson J., Bull V. (2012). A multi-strategy approach to biological named entity recognition. *Expert Systems with Applications, Vol. 39, No. 17*.

Balog K., Ramampiaro H., Takhirov N., Nørvåg K. (2013). Multi-step classification approaches to cumulative citation recommendation. In, p. 121–128. Paris, France. Consulté sur http://

dl.acm.org/citation.cfm?id=2491748.2491775

Baxendale P. B. (1958, octobre). Machine-made index for technical literature: An experiment. *IBM J. Res. Dev.*, vol. 2, n° 4, p. 354–361. Consulté sur http://dx.doi.org/10.1147/rd.24.0354

Berendsen R., Meij E., Odijk M., Daan de Rijke, Weerkamp W. (2012). The university of amsterdam at trec 2012. *Proceedings of The 21th Text Retrieval Conference (TREC)*.

Bikel D., Miller S., Schwartz R., Weischedel R. (1997). Nymble: a high-performance learning name-finder. *Proc. Conference on Applied Natural Language Processing*.

Blei D. M., Ng A. Y., Jordan M. I. (2003, mars). Latent dirichlet allocation. *J. Mach. Learn. Res.*, vol. 3, p. 993–1022. Consulté sur http://dl.acm.org/citation.cfm?id=944919.944937

Breiman L. (2001). Random forests. *Machine Learning, Vol. 45 No. 1*.

Cancedda N., Goutte C., Renders J.-M., Cesa-Bianchi N., Conconi A., Li Y. *et al.* (2002). Kernel methods for document filtering. *Proceedings of The 11th TREC*.

Clark J., Gonzalez-Brenes J. (2008). Coreference resolution : Current trends and future directions.

Cucchiarelli A., Velardi P. (2001). Unsupervised named entity recognition using syntactic and semantic contextual evidence. *Computational Linguistics 27:1.123-131*.

Das D., Martins A. (2007). A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU*.

Davis A., Veloso A., Silva A. da, Jr. W. M., Laender A. (2012). Named entity disambiguation in streaming data. *Proceedings of the 50th Annual Meeting of the ACL*.

Del Corso G. M., Gullí A., Romani F. (2005). Ranking a stream of news. *Proceedings of the 14th International Conference on World Wide Web*, p. 97–106. Consulté sur http://doi.acm.org/10.1145/1060745.1060764

Diaz F. (2009). Integration of news content into web results. *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, p. 182–191. Consulté sur http://doi.acm.org/10.1145/1498759.1498825

Díaz-Uriarte1 R., Andrés S. A. de. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*.

Doddington G., Mitchell A., Przybocki M., Ramshaw L., Strassel S., Weischedel R. (2004). The automatic content extraction (ace) program–tasks, data, and evaluation. *LREC*.

Dong A., Zhang R., Kolari P., Bai J., Diaz F., Chang Y. *et al.* (2010). Time is of the essence: Improving recency ranking using twitter data. *Proceedings of the 19th International Conference on World Wide Web*, p. 331–340. Consulté sur http://doi.acm.org/10.1145/1772690.1772725

Edmundson H. P. (1969, avril). New methods in automatic extracting. *Journal of the ACM*, vol. 16, n° 2, p. 264–285. Consulté sur http://doi.acm.org/10.1145/321510.321519

Frank J., Kleiman-Weiner M., Roberts D., Niu F., Zhang C., Ré C. (2012). Building an entity-centric stream filtering test collection for trec 2012. *Proceedings of The 21th Text Retrieval Conference (TREC)*.

Galibert O., Rosset S., Grouin C., Zweigenbaum P., Quintard L. (2011). Structured and extended named entity evaluation in automatic speech transcriptions. *IJCNLP*.

Hofmann T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 50–57. Consulté sur http://doi.acm.org/10.1145/312624.312649

Ji H., Grishman R., Dang H. (2011). Overview of the tac2011 knowledge base population track. *Proceedings of the Fourth TAC*.

Kjersten B., McNamee P. (2012). The hltcoe approach to the trec 2012 kba track. *Proceedings of The 21th Text Retrieval Conference (TREC)*.

Ko Y., Park J., Seo J. (2002). Automatic text categorization using the importance of sentences. *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, p. 1–7. Consulté sur http://dx.doi.org/10.3115/1072228.1072331

Kohavi R., John G. H. (1997). Wrappers for feature subset selection. *Artif. Intell.*, vol. 97, n° 1-2, p. 273–324. Consulté sur http://dx.doi.org/10.1016/S0004-3702(97)00043-X

König A. C., Gamon M., Wu Q. (2009). Click-through prediction for news queries. *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 347–354. Consulté sur http://doi.acm.org/10.1145/1571941.1572002

Kupiec J., Pedersen J., Chen F. (1995). A trainable document summarizer. *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 68–73. Consulté sur http://doi.acm.org/10.1145/215206.215333

Lam-Adesina A. M., Jones G. J. F. (2001). Applying summarization techniques for term selection in relevance feedback. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 1–9. Consulté sur http://doi.acm.org/10.1145/383952.383953

Lee J. (2012). Mining spatio-temporal information on microblogging streams using a density-based online clustering method. *Expert Systems with Applications, Vol. 39 No. 10*.

Liu X., Fang H. (2012). Entity profile based approach in automatic knowledge finding. *Proceedings of The 21th Text Retrieval Conference (TREC)*.

Liu X., Wei F., Zhang S., Zhou M. (2012). Named entity recognition for tweets. *ACM Transactions on Intelligent Systems and Technology, Vol. 9, No. 4*.

McCallum A. (2003). Early results for named entity recognition with conditional random fields, features induction and web-enhanced lexicons. *Proc. Conference on Computational Natural Language Learning*.

Mihalcea R. (2002). Classifier stacking and voting for text filtering. *Proceedings of The 11th TREC*.

Nadeau D., Sekine S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes, Vol. 30, No. 1*.

Pasca M., Lin D., Bigham J., Lifchits A., Jain A. (2006). Organizing and searching the world wide web of facts-step one: The one-million fact extraction challenge. *Proc. National Conference on Artificial Intelligence*.

Qi X., Davison B. D. (2009). Web page classification: Features and algorithms. *ACM Comput. Surv.*, vol. 41, n° 2, p. 12:1–12:31. Consulté sur http://doi.acm.org/10.1145/1459352.1459357

Radinsky K., Diaz F., Dumais S., Shokouhi M., Dong A., Chang Y. (2013). Temporal web dynamics and its application to information retrieval. *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, p. 781–782. Consulté sur http://doi.acm.org/10.1145/2433396.2433500

Rattenbury T., Good N., Naaman M. (2007). Towards automatic extraction of event and place semantics from flickr tags. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 103–110. Consulté sur http://doi.acm.org/10.1145/1277741.1277762

Robertson S., Soboroff I. (2002). The trec 2002 filtering track report. *Proceedings of The 11th Text Retrieval Conference (TREC)*.

Robertson S., Walker S., Zaragoza H., Herbrich R. (2002). Microsoft cambridge at trec 2002: Filtering track. *Proceedings of The 11th Text Retrieval Conference (TREC)*.

Robertson S. E., Jones K. S. (1997). *Simple, proven approaches to text retrieval*. Rapport technique.

Sebastiani F. (2002, mars). Machine learning in automated text categorization. *ACM Comput. Surv.*, vol. 34, n° 1, p. 1–47. Consulté sur http://doi.acm.org/10.1145/505282.505283

Sekine S. (1998). Nyu: Description of the japanese ne system used for met-2. *Proc. Message Understanding Conference.*.

Sekine S., Nobata C. (2004). Definition, dictionaries and tagger for extended named entity hierarchy. *Proc. Conference on Language Resources and Evaluation*.

Shen D., Chen Z., Yang Q., Zeng H.-J., Zhang B., Lu Y. *et al.* (2004). Web-page classification through summarization. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 242–249. Consulté sur http://doi.acm.org/10.1145/1008992.1009035

Strobl C., Boulesteix A.-L., Kneib T., Augustin T., Zeileis A. (2008, juillet). Conditional variable importance for random forests. *BMC Bioinformatics*, vol. 9, n° 1, p. 307.

Svetnik V., Liaw A., Tong C., Wang T. (2004). Application of breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. *Multiple Classifier Systems, Lecture Notes in Computer Science Volume 3077, pp 334-343*.

Teufel S., Moens M. (1997). Sentence extraction as a classification task. *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scallable Text Summarization*, p. 58–65.

Voorhees E. M. (1999). The trec-8 question answering track report. *NIST Special Publication 500-246: The Eighth Text REtrieval Conference (TREC-8)*.

Yang Y., Pedersen J. O. (1997). A comparative study on feature selection in text categorization. *Proceedings of the Fourteenth International Conference on Machine Learning*, p. 412–420. Consulté sur http://dl.acm.org/citation.cfm?id=645526.657137

Yom-Tov E., Diaz F. (2011). Out of sight, not out of mind: on the effect of social and physical detachment on information need. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, p. 385–394. Consulté sur http://doi.acm.org/10.1145/2009916.2009970