TALN 2010, Montréal, 19–23 juillet 2010

Une approche cognitive de la fouille de grandes collections de documents

Adil El Ghali[1] Yann Vigile Hoareau[1, 2]
(1) Lutin User Lab, Cité des Sciences, Av C. Cariou, 75019 Paris
(2) Université Paris 8, rue de la Liberté, 93200 Saint Denis
elghali@lutin-userlab.fr, hoareau@lutin-userlab.fr

Résumé. La récente éclosion du Web2.0 engendre un accroissement considérable de volumes textuels et intensifie ainsi l'importance d'une réflexion sur l'exploitation des connaissances à partir de grandes collections de documents. Dans cet article, nous présentons une approche de rechercher d'information qui s'inspire des certaines recherches issues de la psychologie cognitive pour la fouille de larges collections de documents. Nous utilisons un document comme requête permettant de récupérer des informations à partir d'une collection représentée dans un espace sémantique. Nous définissons les notions d'identité sémantique et de pollution sémantique dans un espace de documents. Nous illustrons notre approche par la description d'un système appelé BRAT (Blogosphere Random Analysis using Texts) basé sur les notions préalablement introduites d'identité et de pollution sématique appliquées à une tâche d'identification des actualités dans la blogosphère mondiale lors du concours TREC'09. Les premiers résultats produits sont tout à fait encourageant et indiquent les pistes des recherches à mettre en oeuvre afin d'améliorer les performances de BRAT.

Abstract. Mining Web 2.0 content become nowadays an important task in Information Retrieval and Search communities. The work related in this paper present an original approach of blogs mining, inspired from researches in cognitive psychology. We define the notions of semantic identity of blogs, and the semantic pollution in a semantic space. Then, we describe a system called BRAT (Blogosphere Random Analysis using Texts) based on these notions that has been applied to the Top Stories identification task of the Blog Track at the TREC'09 contest. The performance of BRAT at TREC'09 in its preliminary stage of development are very encouraging and the results of the experiences described here-after draw the lines of the future researches that should be realized in order to upgrade its performances.

Mots-clés : Fouille de textes, Random-Indexing, Cognition, Marche aléatoire.

Keywords: Text-Mining, Random-Indexing, Cognition, Random walk.

1 Introduction

Dans le présent article, nous nous intéressons à la recherche d'informations dans de grandes collections de documents en utilisant un ou plusieurs documents comme requête. Nous définissons un système à deux modules pour réaliser cette tâche. Le premier module distribue et représente les documents textuels dans des espaces sémantiques construits avec la méthode Random Indexing (RI) (Kanerva *et al.*, 2000). Le

deuxième module réalise la recherche des documents en utilisant une marche aléatoire [1] pour parcourir l'espace sémantique et trouver les éléments en rapport avec une requête donnée, composée d'un ou plusieurs documents. Le système a été construit en s'appuyant sur deux hypothèses de travail que nous considérons importantes lorsqu'il est question de traiter la sémantique de grandes collections de documents : les notions d'identité sémantique et de pollution sémantique.

L'article est organisé comme suit. Dans la première partie, nous présentons brièvement les modèles d'espaces sémantiques utilisés ainsi que leurs propriétés. Nous définissons ensuite les notions d'identité sémantique et de pollution sémantique qui donnent quelques unes des propriétés principales de l'espace sémantique. Nous décrivons également dans cette partie comment utiliser des documents en tant que requêtes. Dans la deuxième partie, nous présentons une instanciation de notre système conçu pour aborder la tâche d'identification des dépêches d'actualité [2] dans le cadre du TREC'09, nommé BRAT (Blogosphere Random Analysis using Texts). La troisième partie décrit les propriétés et les performances des différentes exécutions soumises dans le cadre du TREC'09.

2 La cognition de la fouille de textes

2.1 Les espaces sémantiques

Les modèles de représentation vectorielle de la sémantique des mots sont une famille de modèles qui représentent la similarité sémantique entre les mots en fonction de l'environnement textuel dans lequel ces mots apparaissent. La distribution de co-occurrence des mots dans le corpus est rassemblée, analysée puis transformée en espace sémantique dans lequel les mots sont représentés comme des vecteurs dans un espace vectoriel de grande dimension. LSA (Landauer & Dumais, 1997), HAL (Lund & Burgess, 1996) et RI (Kanerva *et al.*, 2000) en sont quelques exemples. Ces modèles sont basés sur l'hypothèse distributionnelle de (Harris, 1968) qui affirme que les mots qui apparaissent dans des contextes similaires ont un sens similaire. La caractérisation de l'unité de contexte est une problèmatique commune à toutes ces méthodes, sa définition est différente suivant les modèles. Par exemple, LSA construit une matrice mot-document dans laquelle chaque cellule a_{ij} contient la fréquence d'un mot i dans une unité de contexte j. HAL définit une fenêtre flottante de n mots qui parcourt chaque mot du corpus, puis construit une matrice mot-mot dans laquelle chaque cellule a_{ij} contient la fréquence à laquelle un mot i co-occure avec un mot j dans la fenêtre précédemment définie. Différentes méthodes mathématiques et statistiques permettant d'extraire la signification des concepts, en réduisant la dimensionnalité de l'espace de co-occurence, sont appliquées à la distribution des fréquences stockées dans la matrice mot-document ou mot-mot. Le premier objectif de ces traitements mathématiques est d'extraire les «patrons» qui rendent compte des variations de fréquences et qui permettent d'éliminer ce qui peut être considéré comme du « bruit ». LSA emploie une méthode générale de décomposition linéaire d'une matrice en composantes indépendantes : la décomposition de valeur singulière (SVD). Dans HAL la dimension de l'espace est réduite en maintenant un nombre restreint de composantes principales de la matrice de co-occurrence. À la fin de ce processus de réduction de dimensionnalité, la similitude entre deux mots peut être calculée selon différentes méthodes. Classiquement, la valeur du cosinus de l'angle entre deux vecteurs correspondant à deux mots ou à deux groupes de mots est calculée afin d'approximer leur similarité sémantique.

1. *Random Walk*
2. *Top-stories identification task*

2.2 La notion d'identité sémantique

Lors de l'utilisation de méthodes d'espaces sémantiques pour représenter le sens, nous avons commencé à exposer ci-dessus la l'idée selon laquelle la sémantique produite pour un mot donné dépend de la distribution des autres mots avec lesquels il co-occure. Par conséquent, quelque soit le mot, aucune sémantique ne peut être produite *ex nihilo*, c'est-à-dire sans la réalisation d'une phase d'apprentissage appliquée à une distribution de contextes ou d'épisodes donnée. La sémantique finale associée à un mot dispose ainsi d'une identité forgée tout au long de la phase d'apprentissage, qui est réalisée par la SVD pour LSA ou par le procesus d'*accumulation* pour RI. L'identité sémantique d'un mot donné change en fonction du corpus dans lequel on le retrouve. En prenant l'exemple concret des espaces sémantiques de l'université du Colorado Boulder [3] disponibles en libre accès. Les cinq plus proches voisins du mot «table» dans le «Biology HS betatest » (articles scientifiques) sont : *table* (0.98), *listed* (0.80), *summarized* (0.68), *detail* (0.47), *following* (0.46) [4]. En revanche, le corpus du «General Reading up to 1st year of college » (manuels de lecture) donne comme voisins : *table* (0.98), *tables* (0.68), *centerpiece* (0.65), *dinnerware* (0.62), *tablecloth* (0.61) [5]. Il est aisé de s'apercevoir que les identités sémantiques du mot «table» dans les deux corpus sont très différentes.

La notion d'identité sémantique s'applique non seulement à l'échelle des mots mais également l'échelle de l'espace sémantique. Un espace sémantique dispose d'une identité spécifique qui lui est donné par la distribution de la co-occurrence des mots dont il est composé.

2.3 La notion de pollution sémantique

La notion d'identité sémantique n'est pas révolutionnaire pour des chercheurs familiers avec les espaces sémantiques et pourrait sembler quelque peu triviale si elle ne faisait pas ressortir une seconde notion que nous appellerons « pollution sémantique». Dans le précédent exemple d'un espace sémantique mixte, scientifique et général, l'identité sémantique du mot «table» est constituée autant par la sémantique relative à la science que par celle de la vie quotidienne. Dans un espace sémantique général, si un mot est similaire au mot «table», l'on peut supposer que ce mot n'est pas très éloigné du mot «vaisselle» dans l'espace sémantique. Dans un espace sémantique mixte, une telle supposition paraît moins raisonnable, car la sémantique du mot «table» a été d'une certaine manière «polluée» par la partie scientifique du corpus.

On pourrait soutenir que la pollution sémantique n'est autre que de la polysémie. Cela est vrai dans le cas du mot «table» parce que c'est un mot polysémique, toutefois la pollution de l'identité du mot «table» a pour effet de polluer l'identité de mots sémantiquement proches du mot «table» tels que «récapitulé», «énuméré», «vaisselle», «maison», etc. Ces mots ne sont pas des mots polysémiques mais leurs identités sémantiques se retrouveront polluées aussi. À cause du mot «table», des mots tels que «récapitulé» peuvent être éventuellement semblables au mot «vaisselle» dans un espace sémantique mixte (Beyer *et al.*, 1999; Giannella, 2009). En conclusion, la pollution sémantique adresse non seulement l'échelle du mot mais également l'échelle de l'espace sémantique et de sa structure.

3. http ://lsa.colorado.edu/
4. «table» (0,98), «énuméré» (0,80), «récapitulé» (0,68), «détail» (0,47), «suivant» (0,46)
5. «table» (0,98), «tables» (0,68), «centre de table» (0,65), «vaisselle» (0,62), «nappe» (0,61)

2.4 Des documents comme requêtes

Dans de nombreuses applications, l'utilisateur qui recherche des documents dans une grande collection dispose d'un ou plusieurs documents qui peuvent être utilisés comme requête. Dans l'application décrite dans la section 4 : la tâche d'identification des dépêches importantes [6] dans le cadre de la Blog-Track du TREC'09, l'utilisateur cherche à extraire les *posts* de blogs qui sont pertinentes étant donné une dépêche d'actualité. Dans ce cas, la dépêches peut être utilisé comme requête pour rechercher les éléments pertinants dans l'espace sémantique représentant la blogosphère. Afin d'être en mesure d'utiliser ces documents en tant que requête, nous avons besoin pour représenter documents dans l'espace de recherche. Dans un espace sémantique, les mots sont représentés par des vecteurs dans un espace de dimension n. Cette représentation peut être étendu à documents en associant à chaque document le vecteur correspondant à la somme des mots qu'il contient (Rehder *et al.*, 1998). Le vecteur d'un document est ainsi donné par :

$$\vec{v}_d = \sum_{w \in d} \vec{v}_w \tag{1}$$

Cette représentation permet de réaliser les mêmes calculs sur les documents que sur les mots dans l'espace sémantique. De plus, l'espace les documents composés par les vecteurs des documents a les mêmes propriétés que l'espace les mots. Les notions de l'identité sémantique et de la pollution sémantique sont naturellement étendues aux documents.

3 BRAT

BRAT représente dans un même espace sémantique la production de la blogosphère et la production journalistiques afin de permettre d'identifier pour chacune des dépêches, les posts de blogs pertinents. Il construit des représentations l'identité sémantique liée à l'actualité du point de vue de la blogosphère et du point de vue de la production journalistique, puis établit la pertinence des posts en fonction (i) de leur similarité sémantique avec la dépêche, (ii) de la proximité des posts par rapport à l'identité sémantique du point de vue de la blogosphère et (iii) de la proximité des posts par rapport à l'identité sémantique du point de vue de la production journalistique.

3.1 Construction des espaces sémantiques

La méthode de construction d'espace sémantique utilisé est Random Indexing (RI), qui est relativement éloignée des autres méthodes de construction d'espaces sémantiques. Ses particularités sont (i) qu'elle ne construit pas de matrice de co-occurrence et (ii) qu'elle ne nécessite pas, contrairement aux autres modèles vectoriels de représentation sémantique, des traitements statistiques lourds comme la SVD pour LSA. RI est basée sur la projection aléatoire (Bingham & Mannila, 2001), qui permet un meilleur passage à l'échelle pour grand nombre des documents. La construction d'un espace sémantique avec RI se déroule comme suit :

– Créer une matrice $A(d \times n)$, contenant des vecteurs indexes, où d est le nombre de documents ou de contextes et n le nombre de dimensions choisies par l'expérimentateur. Les vecteurs indexes sont des vecteurs creux générés aléatoirement.

6. *Top-stories identification task*

– Créer une matrice $B(t \times n)$, contenant des vecteurs termes, où t est le nombre de termes différents dans le corpus. Initialiser tous ces vecteurs avec des valeurs nulles pour démarrer la construction de l'espace sémantique.
– Pour tout document du corpus. Chaque fois qu'un terme τ apparaît dans un document δ, accumuler le vecteur index de δ au vecteur terme de τ .

à la fin du processus, les vecteurs termes qui apparaîssent dans des contextes similaires ont accumulé des vecteurs indexes similaire.

RI a été appliqué avec succès au test de synonymie du TOEFL (Kanerva *et al.*, 2000) ainsi que dans la catégorisation de textes (Sahlgren & Cöster, 2004; Hoareau *et al.*, 2009b,a; El Ghali *et al.*, 2009)

Dans notre application, pour chaque date D un espace sémantique SS_D est construit en utilisant la librairie Semantic Vectors [7] (Widdows & Ferraro, 2008). Celui-ci contient toutes les dépêches de la journée et tous les posts dans une fenêtre $[D-1, D+1]$.

3.2 Un *random walk* dans les espaces sémantiques

Une fois l'espace sémantique SS_D d'une journée D construit, nous utilisons un algorithme de marche aléatoire pour naviguer dans l'espace afin de récupérer pour chaque titre n posts pertinents.

Nous appellons prototype d'un ensemble de documents d'une catégorie (posts de blogs ou dépêches), c'est un pseudo-document représenté dans l'espace sémantique par la somme de tous les vecteurs de l'ensemble. Par exemple, le prototype de toutes les dépêches est un document pseudo P_H représentée par le vecteur :

$$\vec{P_H} = \sum_{h \in H} \vec{h} \tag{2}$$

avec H l'ensemble contenant toutes les dépêches de SS_D.

Étant donné une dépêche $h_i \in SS_D$ et $\eta \in N$, nous appelons η-voisinage de h_i pour un prototype P, l'ensemble des posts défini comme suit :

$$\eta-\text{voisinage}(h_i, P) = \{b_j | d(b_j, h_i) < \frac{d(P, h_i)}{\eta}\} \tag{3}$$

avec $d(d_i, d_j)$ la distance euclidienne entre les vecteurs $\vec{d_i}$ and $\vec{d_j}$.

Afin de récupérer les n posts pertinents pour la dépêche h_i, nous choisissons un seuil $m > n$, et nous parcourons aléatoirement l'ensemble $\mathbb{B}$ contenant tous les posts de SS_D, jusqu'à trouver m posts candidats dans l'η-voisinage de h_i pour le prototype P_H de toutes les dépêches. Si nous avons trouvé m posts candidats, nous définissons le score p_i de h_i comme le nombre de pas effectués dans $\mathbb{B}$. Si le nombre de posts récupérés est $m' < m$, alors le score p_i de t_i est définie par :

$$p_i = card(B) - m' \tag{4}$$

Le première application de BRAT dans le cadre du TREC'09 vise essentiellement à tester l'effet du paramètre η-voisinage. Une description en sera donnée dans la section suivante.

7. `http://code.google.com/p/semanticvectors/`

4 Application à la fouille de Blogs

Le modèle BRAT a été appliqué à la tâche d'identification des dépêches dans le cadre du TREC'09. Les objectifs sont de détecter, pour une journée de publication, les dépêches importantes du NewYork Times (NYT) à partir de l'analyse de la blogosphère et pour chacune des dépêches considérée comme importante, de proposer une dizaine de blogs pertinents.

L'expérience rapportée ci-après est le résultat de la participation à un concours, elle n'a donc pas pour objectif d'évaluer systématiquement l'effet de chacun des paramètres du modèle. Mais plutôt à vérifier quelques hypothèses et à éclairer des pistes de travail pour les recherches futures.

4.1 Les hypothèses

Pour chaque journée de publication un espace sémantique est construit à partir des posts produits dans une fenêtre de trois jours et des titres du NYT pour la journée. Une représentation de l'identité sémantique de la blogosphère pour cette fenêtre est approximée en créant un prototype P_B contenant tous les posts de la fenêtre. Une représentation de l'identité sémantique sémantique correspondant à «la vision» de l'actualité par le NYT est approximée un prototype P_H (*cf.* 3.2).

Les exécutions soumises correspondent à différentes hypothèses concernant l'organisation de la connaissance dans l'espace sémantique construit à partir de la blogosphère. Les hypothèses qui ont guidé ce travail sont les suivantes : (i) Le voisinage de P_B est dense et pollué. Il est donc préférable de recruter des blogs qui ne sont pas dans ce voisinage. De plus, différents facteurs devraient améliorer la précision de BRAT : (ii) l'augmentation de la contrainte sur la proximité du voisinage de P_H pour le recrutement des blogs (plus η est petit) ; (iii) la fusion des résultats produits à partir de l'augmentation successives de la contrainte sur le voisinage au prototype (méthode d'adaptative) ; (iv) La combinaison de la contrainte sur le voisinage de P_B et sur le voisinage de P_H ; (v) L'augmentation de la dimension de l'espace sémantique.

Ces hypothèses ont été implantés dans les différentes exécutions décrites dans la sous-section 4.4.

4.2 La collection

La collection Blog08 (cf. Table 1) a été consrtuite par l'aspiration de la blogosphère pendant plus d'un an. Elle est composée de trois éléments : *feeds* correspondant à l'ensemble des résumé des posts d'un blog, *permalinks* correspondant aux posts de blogs et *homepages* contenant les pages d'accueil des auteurs.

La collection	# d'éléments	Taille
Feeds	1.303.520	808GB
Permalinks	28.488.766	1445GB
Homepages	1.011.733	56GB

TABLE 1 – Caractéristiques de la collection Blogs08

Dans le cadre de nos expériences, seuls les documents Permalink (*i.e.* les posts de blogs) ont été utilisés. Les espaces sémantiques produits pour chaque journées sont composés en moyenne de plus de 150 000 posts pour une centaine de titres.

4.3 L'évaluation

4.3.1 Description de la méthode

Les exécutions sont regroupées par les organisateurs puis redistribuées de telle façon à ce que les participants évaluent la validité des exécutions les uns des autres. L'évaluation manuelle se déroule en plusieurs phases.

Dans la première phase, l'ensemble des titres proposés par l'ensemble des équipes sont soumis aux évaluateurs afin qu'ils déterminent les titres qui leur paraissent les plus importants pour la journée. Il est demandé aux évaluateurs de se mettre à la place d'un responsable de l'édition d'un journal et de décider, en fonction de leurs connaissances générales ou d'autres ressources (internet notamment), quels titres devraient être retenus comme important pour la journée. À partir des jugements sur l'importance des titres, une première évaluation des systèmes est réalisée.

Dans la deuxième phase, les jugements de l'importance des titres servent à établir la pertinence des blogs proposés par les systèmes. Les organisateurs regroupent les blogs associés aux titres ayant été jugés "importants" lors de la première phase et les distribuent à l'ensemble des participants pour une évaluation croisée. Pour chaque dépêche, les évaluateurs doivent juger de la pertinence des blogs qui lui ont été associés. La consigne par défaut est de considérer comme pertinent, tout ce qui est en lien avec la dépêche. Autrement dit, un blog qui partage un lien, même ténu, avec la dépêche est considérer comme "pertinent".

4.3.2 Décalage entre la définition de la tâche et la consigne d'évaluation

L'évaluation telle qu'elle a été conduite s'appuye sur l'avis des participants-évaluateurs, à qui il a été demandé de se positionner en tant que rédacteur-en-chef, pour décider de l'importance des titres [8] :

> In essence, you should think like the editor of a newspaper or news website. For each headline, make a decision about whether the headline actually occurred on the query day, and whether you would have placed it on the front page of your news website or newspaper on that day

alors que l'objectif de la tâche était explicitement d'indiquer les titres qui pourrait être jugés importants en tenant compte de ce que les blogeurs produisaient :

> For a given unit of time (e.g. date), systems will be asked to identify the top news stories (similar to what is displayed on the main page of Google Blog Search or Google News), and provide a list of relevant blog posts discussing each news story.

Il y a donc un décalage entre la consigne de la tâche qui était d'identifier les titres pertinents au regard de ce qui est produit par la blogosphère et la consigne de l'évaluation qui est de se mettre à la place d'un responsable de l'édition.

Il est important de noter que l'évaluation des blogs est dépendante de l'évaluation de l'importance des titres ; augmentant ainsi l'effet du décalage sur l'importance des dépêches pour une journée à l'évaluation des blogs étant associés à ces titres. Ainsi, les blogs associés à des titres ayant été jugés non-importants

8. http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG

dans la première phase de l'évaluation (du "point de vue de l'éditeur") ne sont jamais concernés par la deuxième phase d'évaluation portant sur les blogs.

Ce décalage entre la consigne de la tâche et la consigne d'évaluation n'est pas sans poser problème. Comme l'ont montré (Balog *et al.*, 2009), pour des jours comme ceux de la fête des mères ou les lendemains de compétitions, les échanges sur la blogosphère sont très fortement impactés par ces thèmatiques. Or, il est tout à fait probable qu'un événement majeur tel qu'un séisme ou qu'un scandale politique se produise pour ces mêmes jours. L'éditeur (et donc l'évaluateur) les considérera comme important tandis qu'un grand nombre de blogeurs souhaiteront de joyeuses fêtes à leur mère ou discuteront des performances de leur équipe favorite.

Ainsi, ce décalage, tout à fait dépendant des choix des organisateurs, est de nature à contrarier l'appréciation sereine des performances des équipes et des systèmes.

4.4 Description des exécutions

Les exécutions décrites ici correspondent aux différentes hypothèses de travail décrites dans la sous-section 4.1. Les hypothèses sur la constraintes de la restriction du voisinage au prototype des dépêches P_H et au prototype des blogs P_B sont implantées en fonction des valeurs de η qui leur est associé.

Ainsi, *ri1025rw2b* correspond à une exécution qui choisit les blogs avec la seule contrainte qu'il soit à une certaine distance par rapport au voisinage du prototype des blogs P_B, avec $\eta = 2$. L'exécution *ri1025rw5432* correspond à un algorithme adaptatif utilisant le même principe, et où les résultats de la marche aléatoire avec η = 5, 4, 3, 2 sont combinés. L'exécution *ri1025rw5h2b* utilise un algorithme similaire, mais utilisant une fonction de voisinage correspondant à l'intersection du 5-voisinage par rapport à P_T et du 2-voisinage par rapport à P_B. Tandis que les précédentes exécutions utilisaient un espace sémantique à 1025 dimensions, l'exécution *ri2049rw3* correspond à une application de l'algorithme dans un espace de 2049 dimensions en prenant pour unique contrainte la distance au voisinage du prototype des dépêches P_H avec η = 3.

4.5 Résultats

Les résultats qui sont présentés ci-après correspondent aux performances pour l'identification des titres importants pour une journée donnée. La mesure utilisées et R-précision@10 qui correspond à la valeur de précision calculée pour 10 dépêches correctement identifiées. Les différentes caractéristiques des exécutions décrites ci-après ainsi que leur performances sont rappelées dans le Tableau 2.

Parmi les différentes méthodes testés, le nombre de dimension de l'espace sur lequel est appliqué la marche aléatoire semble jouer un rôle déterminant. Ainsi, l'exécution *ri2049rw3* donne les meilleurs résultats. Par ailleurs, la méthode adaptative qui combine les résultats de plusieurs marches aléatoires donne de meilleurs résultats que la méthode qui réalise la marche aléatoire en prenant comme double contraintes les voisinage à P_T et à P_B (respectivement *ri1025rw5432* et *ri1025rw5h2b*) . Enfin, la seule contrainte sur le voisinage de P_B donne les moins bonnes performances (*ri1025rw2b*).

	Paramètres de Brat			Performance	
Exécution	Dimension de RI	Valeur de η pour		R-Precision@10	Position
		P_T	P_B		
ri1025rw2b	1025	x	2	0,0964	< Md
ri1025rw5h2b	1025	5	2	0,1145	< Md
ri1025rw5432	1025	ada{5-2}	x	0,1200	> Md
ri2049rw3	2049	3	x	0,1182	> Md

TABLE 2 – Paramètres des exécutions, valeurs des R-Precision@10 et positionnement par rapport à la valeur médiane (Md) sur l'ensemble de participants à la Blog-Track.

5 Conclusion

Nous avons décrit une approche qui permet la recherche d'information dans des grandes collections de documents représentées dans un espace sémantique et interrogées en utilisant une marche aléatoire. L'approche consiste à construire avec RI un espace sémantique pour une journée de publication à partir des posts de la blogosphère et des titres édités par un journal pour le même jour afin d'identifier les titres les plus importants.

BRAT construit une représentation de l'actualité du point de vue de la blogosphère et du point de vue du journal et évalue la pertinence des titres et des blogs associés en prenant en compte leurs similarité sémantique et leur proximité avec la représentation de l'identité sémantique de l'actualité du point de vue de la blogosphère ainsi qu'avec la représentation de l'identité sémantique de l'actualité du point de vue du journal. BRAT réalise une marche aléatoire dans l'espace sémantique et détermine l'importance d'un titre en évaluant le nombre de blogs qui lui sont similaires étant donnée certaines contraintes dépendantes de la marche. Les expériences réalisées ont montrées que les contraintes sur le voisinage lié à représentation de l'actualité du point de vue du journal est plus efficace que les contraintes sur le voisinage liée à la représentation de l'actualité du point de vue de la blogosphère et que la conjugaison de contraintes sur le voisinage liés au deux types de représentations améliore les résultats, de même que l'utilisation de méthodes adaptatives ou l'augmentation de la dimension de l'espace construit avec RI.

Ces différents éléments constituent autant de pistes de recherche qui nous permettrons d'améliorer notre modèle.

Remerciements

Cette recherche a bénéficié de l'aide généreuse et solidaire des sociétés Thalès et Pertimn ainsi que du Pôle de Compétitivité Cap-Digital de la Région Ile de France. Nous leur adressons nos sincères remerciements. Nous remercions par ailleurs tous les collègues du Lutin UserLab.

Références

BALOG K., BRON M., HE J., HOFMANN K., MEIJ E. J., DE RIJKE M., TSAGKIAS E. & WEERKAMP

W. (2009). The University of Amsterdam at TREC 2009 : Blog, web, entity, and relevance feedback. In *TREC 2009 Working Notes* : NIST.

BEYER K. S., GOLDSTEIN J., RAMAKRISHNAN R. & SHAFT U. (1999). When is "nearest neighbor" meaningful ? In C. BEERI & P. BUNEMAN, Eds., *ICDT*, volume 1540 of *Lecture Notes in Computer Science*, p. 217–235 : Springer.

BINGHAM E. & MANNILA H. (2001). Random projection in dimensionality reduction : Applications to image and text data. In *in Knowledge Discovery and Data Mining*, p. 245–250 : ACM Press.

EL GHALI A., HOAREAU Y. & EL GHALI K. (2009). The Episodic Memory Metaphor for Opinion Judgment Categorization. In *IADIS International Conference WWW/Internet (2)*, Rome.

GIANNELLA C. (2009). New instability results for high-dimensional nearest neighbor search. *Inf. Process. Lett.*, **109**(19), 1109–1113.

HARRIS Z. (1968). *Mathematical Structures of Language*. New York : John Wiley and Son.

HOAREAU Y., EL GHALI A., LEGROS D. & EL GHALI K. (2009a). Random Indexing and the episodic memory metaphor. Application to text categorization. In A. ABDOLLAHZADEH & H. PEDRAM, Eds., *IEEE CSICC 2009. 14th International CSI*, Teheran, Iran : IEEE.

HOAREAU Y. V., EL GHALI A. & TIJUS C. (2009b). Detection of opinions and facts. A cognitive approach. In *Recent Advance in Natural Language Processing (RANLP'09)*, Borovets, Bulgaria.

KANERVA P., KRISTOFERSON J. & HOLST A. (2000). Random Indexing of Text Samples for Latent Semantic Analysis. In L. GLEITMAN & A. JOSH, Eds., *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, Mahwah : Lawrence Erlbaum Associates.

LANDAUER T. K. & DUMAIS S. T. (1997). A Solution to Plato's Problem : The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review*, **104**(2), 211–240.

LUND K. & BURGESS C. (1996). Producing high-dimensional semantic space from lexical co-occurence. *Behavior research methods, instruments & computers*, **28**(2), 203–208.

REHDER B., SCHREINER M., WOLFE M., LAHAM D., LANDAUER T. & KINTSCH W. (1998). Using Latent Semantic Analysis to assess knowledge : Some technical considerations. *Discourse Processes*, **25**(2), 337–354.

SAHLGREN M. & CÖSTER R. (2004). Using bag-of-concepts to improve the performance of support vector machines in text categorization. In *COLING '04 : Proceedings of the 20th international conference on Computational Linguistics*, p. 487, Morristown, NJ, USA : Association for Computational Linguistics.

WIDDOWS D. & FERRARO K. (2008). Semantic Vectors : A Scalable Open Source Package and Online Technology Management Application. In *Proceeding of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.