Algorithme automatique non supervisé pour le Deft 2012

Murat Ahat [1] Coralie Petermann [1, 2] Yann Vigile Hoareau [3] Soufian Ben Amor [1] Marc Bui [2]

(1) Prism, Université de Versailles Saint-Quentin-en-Yvelines, 35 avenue des Etats-Unis, F-78035 Versailles.
(2) LaISC, Ecole Pratique des Hautes Etudes, 41 rue Gay-Lussac, F-75005 Paris.
(3) CHArt, 41 rue Gay-Lussac, F-75005 Paris.
murat.ahat@prism.uvsq.fr, coralie.petermann@laisc.net,
hoareau@lutin-userlab.fr, soufian.ben-amor@uvsq.fr, marc.bui@ephe.sorbonne.fr

RÉSUMÉ

Nous décrivons l'approche mise en oeuvre dans le cadre du Défi de Fouille de Texte 2012 pour la piste 1 qui consistait à identifier, pour un article scientifique et son résumé donnés, la liste des mots clés qui lui correspondent parmi un ensemble de mot clés possibles. Cette approche est basée sur le couplage entre les méthodes d'espaces sémantiques pour la représentation des connaissances sémantiques d'une part, et les graphes pour la décision sur l'affectation d'un mot clé à un article, d'autre part. La méthode proposée est entièrement automatique, sans phase de paramétrage, non-supervisée et ne nécessite aucune ressource externe.

ABSTRACT

Automatic unsupervised algorithm for Deft 2012

We describe our approach in Deft 2012 for track 1, which consist in identifying a corresponding list of key word, for a given scientific paper and summary, from a set of possible key words. The approach is based on the one hand, semantic space for the representation of semantic knowledge, and, on the other hand, graphs for the decision on the allocation of a key word to a document. The proposed method is fully automatic, without any particular tuning, unsupervised and requires no external resources.

MOTS-CLÉS : Espace sémantique, Graphe, Random Indexing.

KEYWORDS: Semantic Space, Graph, Random Indexing.

JEP-TALN-RECITAL 2012, Atelier DEFT 2012: DÉfi Fouille de Textes, pages 69–75,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Dans cette édition 2012 du Défi Fouille de Texte, nous avons appliqué notre méthode déjà présentée lors du Deft 2011 (Hoareau *et al.*, 2011b), qui consiste à mixer deux méthodes de représentation des connaissances : les espaces sémantiques qui sont des espaces vectoriels à grandes dimensions et les modèles de graphes. L'intérêt du couplage des deux approches est de bénéficier d'une part des propriétés d'apprentissage non-supervisé ainsi que des propriétés sémantiques latentes associés aux espaces sémantiques et, d'autre part de la sophistication des mathématiques sous-jacentes à la théorie des graphes.Pour ce faire, la première contrainte à respecter est de produire un graphe ayant les mêmes propriétés que l'espace sémantique en ce qui concerne la représentation des relations sémantiques latentes entre les mots ou les documents (Louwerse *et al.*, 2006). Cette contrainte satisfaite, des applications peuvent alors être réalisées directement à partir du graphe. Un exemple d'application de cette approche mixte est celui de la visualisation des relations sémantiques latentes entre documents au sein de grandes bases de données textuelles (Hoareau *et al.*, 2011a).

L'an passé, le challenge consistait à associer un article à son résumé. Dans la suite de cet article, nous allons voir si cette année, toujours sans paramétrage ni apprentissage, notre méthode produit d'aussi bon résultats pour la tâche 1 qui consiste à apparier un article et son résumé à une liste de mots clés. Cette méthode a été instanciée de telle sorte à représenter la relation sémantique entre chaque mot clé et chaque couple article/résumé dans un graphe construit à partir d'un espace sémantique, puis à utiliser ce graphe complet pour associer à chaque article un ou plusieurs mot clé.

L'article est organisé de la façon suivante. Dans la première section nous décrivons le cadre théorique de notre algorithme en présentant les espaces sémantiques et les algorithmes de création de tels espaces à partir d'un quelconque contenu, ainsi que les bases de la théorie des graphes nécessaires à notre approche, afin de représenter les documents sous la forme d'un graphe ayant les mêmes propriétés que l'espace sémantique construit. Dans la deuxième section, nous décrivons notre algorithme. Dans le troisième section, nous présentons brièvement les résultats de notre approche et les comparons avec les résultats obtenus l'an passé. Enfin, nous concluons l'article en présentant les perspectives de recherche qui pourraient prolonger le présent travail.

2 Cadre théorique

2.1 Les espaces sémantiques

La théorie des espaces sémantiques est un ensemble de méthodes algébriques permettant de représenter des documents de tout type selon leur contenu. Plusieurs méthodes permettent de modéliser des espaces sémantiques. Elles admettent toutes l'hypothèse distributionnelle suivante : les mots ayant un sens proche apparaissent dans des documents similaires. Mais toutes reposent sur la sémantique vectorielle : les corpus sont analysés et modélisés sous forme de vecteurs à grandes dimensions, rassemblés dans une matrice de co-occurences. Cette matrice peut être construite de deux manières selon les algorithmes :
– matrice mots-documents, utilisée par exemple dans LSA et RI, qui compte le nombre d'occur-

rences de chaque mot dans chaque document
– matrice mots-mots, utilisée par HAL, qui regroupe les probabilités de co-occurences pour chaque groupe de mots

Etant donnée une représentation vectorielle d'un corpus de documents, on peut introduire une notion d'espace vectoriel permettant de mettre en place la notion mathématique de proximité entre documents. En introduisant des mesures de similarité adaptées, on peut quantifier la proximité sémantique entre différents documents. Les mesures de similarité sont choisies en fonction de l'application.

Une mesure très utilisée est la similarité cosinus, qui consiste à quantifier la similarité entre deux documents en calculant le cosinus de l'angle entre leurs vecteurs. Ainsi, un cosinus nul, signe de l'orthogonalité des deux vecteurs, indiquera que ces 2 documents n'ont aucun mot en commun. L'avantage de cette méthode est que la longueur des documents n'influe en rien le résultat obtenu.

Une autre mesure possible est la distance de Manhattan (appelée aussi city-block), qui elle, prend en compte la longueur des documents comparés.

2.2 Random Indexing

Random Indexing (Kanerva *et al.*, 2000) est un modèle d'espace sémantique basé sur des projections aléatoires.

La méthode de construction d'un espace sémantique avec RI est la suivante :
– Créer une matrice A ($d \times N$), contenant des *vecteurs-index*, où d est le nombre de documents ou de contextes correspondant au corpus et N, le nombre de dimensions ($N > 1000$) défini par l'expérimentateur. Les vecteurs-index sont creux et aléatoirement générés. Ils consistent en un petit nombre de (+1) et de (-1) et de centaines de 0 ;
– Créer une matrice B ($M \times N$) contenant les *vecteurs-termes*, où M est le nombre de termes différents dans le corpus. Pour commencer la compilation de l'espace, les valeurs des cellules doivent être initialisées à 0 ;
– Parcourir chaque document du corpus. Chaque fois qu'un terme τ apparaît dans un document d, il faut *accumuler* le vecteur-index correspondant au document d au vecteur-terme correspondant au terme τ.

À la fin du processus, les vecteurs-termes qui sont apparus dans des contextes (ou documents) similaires, auront accumulé des vecteurs-index similaires.

Cette méthode a démontré des performances comparables (Kanerva *et al.*, 2000) et parfois même supérieures (Karlgren et Sahlgren, 2001) à celles de LSA pour le test de synonymie du TOEFL (Landauer et Dumais, 1997). *RI* a été aussi appliqué à la catégorisation d'opinion (Sahlgren et Cöster, 2004).

2.3 Théorie des graphes

La théorie des graphes est une théorie informatique et mathématique. Cette théorie est largement utilisée dans tous les domaines liés à la notion de réseau (réseau social, réseau informatique, télécommunications, etc.) et dans bien d'autres domaines (génétique, transports...).

Un graphe $G = (V,A)$ est une paire composée de (Berge, 1970) :

1. un ensemble $V = \{x_1, x_2, ..., x_n\}$ appelé *sommets* (en référence aux polyèdres) ou *noeuds* (en référence à la loi des noeuds).
2. une famille $A = (a_1, a_2, ..., a_n)$ d'éléments du produit Cartésien $V \times V = \{(x, y)/x \in V, y \in V\}$ appelés *arcs* (cas d'un graphe orienté) ou arêtes (cas d'un graphe non orienté).

En général, on note n le nombre de noeuds (aussi noté $|V(G)|$) et m le nombre d'arcs (aussi noté $|A(G)|$).

Un chemin P est composé de k arcs tels que $P = (a_1, a_2, ..., a_i, ..., a_k)$ où pour chaque arc a_i, la fin coïncide avec le début de a_{i+1}. Une chaine est l'équivalent d'un chemin dans le cadre non orienté.

Un graphe est simple si au plus une arête relie deux sommets et s'il n'y a pas de boucle sur un sommet. Dans les cas où une arête relie un sommet à lui-même (une boucle), ou plusieurs arêtes relient deux mêmes sommets, on appelle ces graphes des multigraphes.

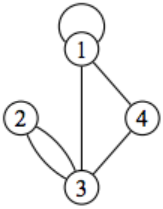
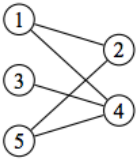


FIGURE 1 – Multigraphe.

Un graphe est biparti si ses sommets peuvent être divisés en deux ensembles X et Y , de sorte que toutes les arêtes du graphe relient un sommet dans X à un sommet dans Y (dans l'exemple ci-dessous, on a X = 1,3,5 et Y = 2,4).



Graphe biparti

$V = \{1,2,3,4,5\}$
$E = \{\{1,2\},\{1,4\},\{2,5\},\{3,4\},\{4,5\}\}$

FIGURE 2 – Graphe biparti.

Dans notre méthode, nous utiliserons des graphes simples bipartis, afin d'associer chaque article à une liste de mots clés. Dans la section suivante, nous présentons notre algorithme en détails.

3 Notre algorithme

Cette section décrit le processus de construction (i) d'un graphe complet représentant les propriétés sémantiques d'un espace sémantique, puis (ii) d'un graphe biparti à partir d'un espace sémantique.

Notre méthode débute par une étape de prétraitements qui consiste à supprimer les mots vides de sens tels que les conjonction de coordination, articles indéfini, pronoms...

Le procédé consiste ensuite à générer notre espace sémantique à l'aide de la méthode RI puis à calculer la distance euclidienne pondérée entre chaque document et chaque mots clés de l'espace sémantique afin de construire un graphe biparti complet. L'intérêt de cette méthode très simple est de générer automatiquement un graphe biparti et de permettre ainsi d'y appliquer les méthodes issues de la théorie des graphes (Hoareau *et al.*, 2011a).

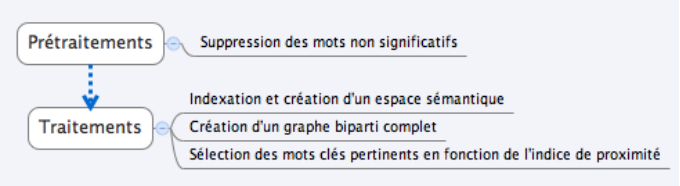


FIGURE 3 – Notre algorithme.

L'algorithme décrit ci-après a pour objectif de construire un graphe biparti à partir d'un espace sémantique. Il prend en entrée un ensemble d'articles . Une matrice m "article – mots clés" est construite. Cette matrice contient dans chaque cellule $m_{i,j}$, la valeur de la distance euclidienne pondérée entre les vecteurs de l'article i et du mots clés j. À partir de cette matrice, un graphe biparti complet g est produit. Un processus de filtre est appliqué à ce graphe afin de produire un graphe biparti où à un article est connecté à ces mots clés.

```
Procedure main()
 Var
   A as Article Set;
   K as Kew word Set;
   N as number of articles;
   M as number of keywords;
   m as Matrix Article Key word;
   g as graph (article --> key word);

 Begin
   spaceSemantic = RandomIndexing(A)

   For (i:=1 to N)
     artVector = spaceSemantic(A[i]);
```

```
      For (j:=1 to M)
        keyVector = spaceSemantic(R[j]);
        m[i,j] = cosine(artVector, resVector);
      End For; //j
    End For; //i

    g = createGraph(m);
  End Procedure //main()


Procedure createGraph(m);
  Var
    m as Matrix Article Key word;
    g as graph (article --> key word);
    knumList as number of keywords for articles;
  Begin
    g = emptyGraph();
    For (i:= 1 to N)
      templist = Max(m[i,:],knumbList);
      g.add(i,templist);
    End
    Return g;
  End Procedure //createGraph()
```

4 Résultats et discussion

Pour le défi Deft 2012, nous avons soumis deux groupes de résultats, obtenus à l'aide de deux espaces sémantiques différents. Le premier est créé à partir des documents de test et d'apprentissage, alors que le second est créé uniquement à partir des documents de test. La librairie utilisée implémentant random indexing est semantic vectors, et la dimension des vecteurs a été paramétrée à 2048 avec un cycle d'entrainement. Même si notre algorithme nous a fourni de bons résultats pour Deft 2011 en nous hissant sur la première place ex aequo du podium (Hoareau *et al.*, 2011b), les résultats obtenus cette année ne sont pas satisfaisants (voir le tableau suivant).

Run	Précision	Rappel	F-score
1	0,0428	0,0428	0,0428
2	0,0242	0,0242	0,0242

TABLE 1 – Scores pour les tâches d'appariement du DEFT 2012

Nous avons tenté en vain d'améliorer nos résultats avec des paramétrages différents des espaces sémantiques pour tester des dimensions jusqu'à 6000 et jusque 5 cycle d'entrainement. Nous assumons alors que la méthode de random indexing peut être une des causes de cet échec. Nous poursuivons donc nos recherches sur ce sujet, en testant diverses méthodes de constructions d'espaces sémantiques et divers outils concernant les espaces sémantiques.

5 Conclusion et perspectives

La méthode proposée dans le cadre de notre participation au Deft repose sur le couplage entre les espaces sémantiques et les graphes. Le faible nombre de documents disponibles pour l'apprentissage constituait une contrainte forte pour notre méthode entièrement basée sur une approche distributionnelle. En 2011, nous avions obtenu de bons résultats mais la tâche du Deft 2012 a montré les limites de notre méthode.

De prochaines expériences seront réalisées afin de comparer notre méthode, et améliorer son paramétrage.

Références

BERGE, C. (1970). *Graphes et Hypergraphes*. Dunod, Paris.

HOAREAU, Y. V., AHAT, M., MEDERNACH, D. et BUI, M. (2011a). Un outil de navigation dans un espace sémantique. *In* KHENCHAF, A. et PONCELET, P., éditeurs : *Extraction et gestion des connaissances (EGC'2011)*, volume RNTI-E-20 de *Revue des Nouvelles Technologies de l'Information*, pages 275–278. Hermann-Éditions.

HOAREAU, Y. V., AHAT, M., PETERMANN, C. et BUI., M. (2011b). Couplage d'espaces sémantiques et de graphes pour le deft 2011 : une approche automatique non supervisée. *In DÉfi Fouille de Textes (DEFT 2011)*, Montpellier, France.

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

KARLGREN, J. et SAHLGREN, M. (2001). From Words to Understanding. *In* UESAKA, Y., KANERVA, P. et ASOH, H., éditeurs : *Foundations of Real-World Intelligence*. CSLI Publications, Stanford.

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

LOUWERSE, M., CAI, Z., HU, X., VENTURA, M. et JEUNIAUX, P. (2006). Cognitively inspired natural-language based knowledge representations : Further explorations of latent semantic analysis. *International Journal of Artificial Intelligence Tools*, 15:1021–1039.

SAHLGREN, M. et 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*, page 487, Morristown, NJ, USA. Association for Computational Linguistics.