LA MODELISATION DU TEXTE EN SIGNAL : VERS UN NOUVEAU MODELE DE REPRESENTATION DE L'INFORMATION

Nabila SMAIL (*) Renaud EPPSTEIN(*)
nsmail@univ-mlv.fr eppstein@univ-mlv.fr

(*) Laboratoire : Sciences et Ingénierie de l'Information et de l'Intelligence Stratégique S3IS ; Université de Marne La Vallée
Cité Descartes 5, bd Descartes
Champs sur Marne
77454 MARNE LA VALLEE
CEDEX 2

Mots clefs :
Systèmes de recherche d'informations, modélisation textuelle, Analyse Multiresolution, mots associés, Ondelettes
Keywords:
Information Retrieval Systems, textual modelling, Multiresolution analysis, co-word analysis, Wavelet transform
Palabras clave:
Sistemas de Recuperación de Información , modelización textual, Análisis de multiresolución, palabras asociadas, Olas pequeñas
Catégorie: Recherche d'information

Défit et but :

Les avancements récents et le développement dans la gestion des réseaux, les télécommunications et les technologies de stockage durant ces dernières années, ont causé une explosion massive dans la quantité d'information disponible. Le public à suivie cette tendance en se connectons sur Internet, ou chacun peut chercher ou éditer sa propre information. Beaucoup de questions peuvent trouvées des réponses dans le Web, mais cela ne signifie pas que nous pouvons toujours les trouver.

Pour traiter toute cette information, l'utilisateur exige l'organisation des données trouvées dans des grandes collections de textes. Par exemple, chaque email peut être considéré comme un texte. Des millions d'email sont envoyés par Internet chaque jour. Chaque email est habituellement enregistré sur l'ordinateur du transmetteur et celui du récepteur. À moins que ces emails soient organisés en catégories par le transmetteur et le récepteur, ils auront du mal à localiser des emails spécifiques si nécessaires.

Le défi actuel consiste à développer un système qui peut stocker de grandes quantités d'information textuelle et les organiser de façon à pouvoir les recherchées facilement.

Ce travail se concentre sur les aspects fondamentaux de traitement de textes dans la recherche documentaire. Nous allons voir que beaucoup de systèmes de recherche ne nous fournissent pas toujours les résultats que nous prévoyons. Nous supposerons que c'est dû au fait que beaucoup de méthodes considèrent seulement la fréquence des mots dans les documents, plutôt que leur position.

Ce travail développe un nouveau paradigme pour représenter des mots dans un document sous forme de signal appelé un signal de thème. Ce signal nous permet de tracer l'information contenue dans le document dans domaine spectral. L'analyse Multi-résolution est ensuite appliquée aux signaux de thèmes pour analyser l'effet de l'application des Ondelettes sur le signal en termes de précision et particulièrement pour la compression.

Résumé

L'évolution de la société de l'information et l'avènement d'Internet ont rendu disponible une grande masse d'information et son volume augmente considérablement chaque année. L'un des problèmes de cette extraordinaire croissance pour l'utilisateur est de trouver l'information pertinente par rapport à un besoin généralement exprimé à l'aide d'une requête.

Pour que les outils de recherche dans les multiples sources d'informations électroniques deviennent un instrument réel de travail, on ne peut se contenter du niveau actuel de performances des modèles de recherche (Un état de l'art en matière de Système de Recherche d'Informations montre une marge d'amélioration considérable).

Dans cet article, nous proposons un modèle de représentation de l'information contenue dans les documents électroniques. Pour ce faire, nous explorons une nouvelle voie dans la représentation de l'information en modélisant le texte sous la forme d'un signal numérique. L'originalité de cette représentation repose sur la façon de considérer un document, non plus simplement comme « un sac de mots » mais comme un ensemble de signaux décrivant la présence et l'importance relative de thématiques choisies. Cette nouvelle forme de représentation documentaire nous permettra, dans la suite de nos travaux, d'appliquer de nombreux outils mathématiques bien connus en théorie du signal, tel que les transformées en ondelettes par exemple, afin de proposer des méthodes pertinentes et innovantes de classification et de comparaison textuelle.

Ce type de représentation de l'information a déjà été évoqué par Miller et al [3] qui l'implémentent dans le cadre de la construction d'un prototype de visualisation nommé Topic-O-Graphy. Ce prototype est utilisé pour la visualisation et l'analyse graphique des ruptures thématiques dans un texte non- structuré. En 2003, Al Halimi [2] développe également cette voie pour l'intégrer dans des applications de classification documentaire.

Dans cet article, nous développons et précisons certains concepts évoqués dans les travaux ci-dessus. De plus, nous proposons une méthode originale de construction du signal se fondant sur la théorie des mots associés. Plus particulièrement, le flot thématique d'un thème choisi dans un document sera tracé et visualisé comme un signal, dont la valeur varie en fonction de la présence d'un terme porteur de sens par rapport à ce thème. Nous illustrons notre propos par un exemple.

Summary

The evolution of the information system and the Internet made available a great mass of information and its volume increases considerably every year. One of the problems of this extraordinary growth, for the user, is to find the relevant information related to a need generally expressed by a request. And because the tools of search in the multiple sources of electronic information become a real instrument of work, we cannot be satisfied with the current level of performances of the Information System models.

In this article, we propose a model of representation of the information contained in the electronic documents. To do this, we explore a new way in the representation of information by modeling the text in the shape of a numerical signal. The originality of this representation consists in the way of considering a

document, not simply as a “bag of words” but as a set of signals describing the presence and the importance of such thematic selected sets.
This new representation will allow us to apply many mathematical tools known in the theory of the signal such as, Wavelets Transformations in order to propose new methods of classification and textual comparison.

This type of representation of information was already proposed by Miller and al. [3] that implemented it within the framework of a prototype of visualization named Topic-O-Graphy. This prototype is used for the visualization and the graphic analysis of the ruptures sets of themes in a text. In 2003, Al Halimi [2] also developed the same way and integrated it in different applications of documentary classification.

In this article, we develop and specify some concepts proposed in the previous work. Moreover, we propose an original method of construction of the signal based on the theory of the co-word analysis. More particularly, the flood of a topic chosen in a document will be traced and visualized as a signal, whose value varies according to the presence of a term carrying direction compared to this topic and we illustrate our method by an example

1 Introduction

La recherche d'information (Information Retrieval) est un vaste domaine de recherche au carrefour des sciences de l'information, de l'informatique, de la linguistique et des mathématiques. L'idée sous-jacente d'un système de recherche d'information (SRI) est que le degré d'appariement entre la requête de l'utilisateur et les documents de la base sur laquelle s'effectue l'interrogation, permet d'indiquer la pertinence des documents retrouvés. L'objectif essentiel de la recherche d'information consiste à améliorer les performances de l'interrogation selon deux mesures : le rappel et la précision.

Il existe un grand nombre de modèles fournissant des représentations différentes de l'information contenue dans les documents. Ces représentations explicitent généralement les relations possibles entre les mots d'un index. Dans l'histoire de la recherche d'information, on constate d'abord l'utilisation large du modèle booléen pour sa simplicité, aussi bien dans les bases bibliographiques que pour les moteurs de recherche. Cependant sa version sans pondération comporte quelques faiblesses. Plusieurs auteurs [14, 15] proposent des extensions de ce modèle : booléen étendu, logique flou.

Le modèle vectoriel, populaire depuis longtemps, se distingue par sa capacité à ordonner les documents trouvés. Les documents et les requêtes sont représentés dans un espace d'information multidimensionnel et la pertinence d'un document par rapport à une requête est à estimé par des mesures définies sur cet espace. Le modèle probabiliste a également montré son efficacité dans la recherche et la comparaison documentaire, en particulier dans les compagnes d'évaluation TREC (Text REtrieval Conferences). Cette approche s'intéresse à la probabilité de pertinence des documents par rapport à la requête.

Aujourd'hui, un système de recherche d'information doit être capable de traiter de très grands corpus et de présenter des résultats fiables dans un temps minimum. Ainsi, si l'on considère la chaîne de traitement de l'information, l'objet de notre article se porte plus précisément sur la modélisation de l'information textuelle pour l'analyse. La proposition de modélisation que nous formulons consiste à représenter un texte à l'aide des signaux numériques, mode de représentation issu de la théorie du signal et généralement utilisé pour le traitement de l'image et du son.

2 Vers une modélisation du texte en signal

2.1 Introduction

La modélisation des documents dans les modèles traditionnels (booléen, vectoriel) considère un document comme un simple « sac de mots ». Cette approche se fonde sur le principe que les mots (considérés individuellement) utilisés dans un document sont largement suffisants pour décrire le thème principal du texte. Elle suppose aussi que la probabilité d'utiliser un mot dans un texte est strictement indépendante des autres mots. En ce sens, un texte est considéré comme une large entité sans structure informationnelle.

Dans le cadre de ce travail, nous proposons une nouvelle approche qui prend en considération le contexte des mots ainsi que leurs ordres d'apparition, pour extraire les mots candidats qui indiquent la présence et la variation thématique dans des longs documents. En effet, la variation thématique dans un texte implique le

changement des mots employés. Tracer l'évolution du discours à travers le texte (marqué par l'apparition des mots dans leurs ordres chronologiques), revient à tracer la courbe représentant le changement thématique à travers le texte en temps réel et à différents niveaux de détails.

Le recours à la modélisation des textes en signaux parait un choix judicieux en considérant les arguments suivants :

- la modélisation en signaux a déjà prouvé son utilité dans diverses applications, notamment le traitement d'images, de la parole et du son.
- une telle approche permet de faire appel à différents outils mathématiques empruntés à la théorie du signal et jusque là inexploités dans un contexte de comparaison textuelle.

2.2 Présentation de la méthode

L'hypothèse que nous formulons est que le texte peut être représenté par un signal représentant la présence et le changement thématique à travers le texte. L'objectif de cette modélisation est l'utilisation ultérieure des méthodes mathématiques à des fins de compression et de comparaison de données.

Nous fondons nos réflexions sur des travaux [1] réalisés au sein du Pacific Northwest National Laboratory en 1998, dans lesquels les auteurs présentent un prototype de visualisation et exploration de données non structurées 'Topic-O-Graphy-. Cette proposition applique une transformée en ondelettes sur un signal numérique correspondant au texte afin de visualiser les ruptures thématiques dans un long document.

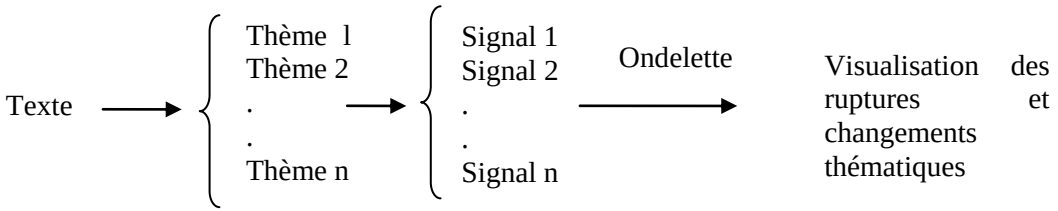

Figure 2. Prototype de visualisation de ruptures thématiques d'après Miller et al [3]

2.2.1 Description

Le processus de modélisation décrit par Miller [3] commence par l'identification des mots porteurs de sens. Pour cela, les auteurs utilisent les mesures de pertinence proposées par Bookstein [4] : Les mots ayant un poids plus élève qu'un seuil fixé sont considérés comme des thèmes. Ceux avec un poids faible, mais supérieur à un second seuil, inférieur, sont appelés des mots croisés. Tous les mots restants avec des poids inférieurs sont rejetés.

Les méthodes de sélection thématique de Bookstein, consistent à appliquer des calculs de comparaison entre l'occurrence du mot dans le texte et son occurrence aléatoire prévue.
Bookstein définit deux mesures statistiques :

- la tendance du mot à se regrouper en masse, dans une unité textuelle étroite (telle qu'un paragraphe).
- la tendance d'un mot de se reproduire au moins une fois dans plusieurs unités textuelles consécutives dans un document.

Ces méthodes de mesure montrent qu'une telle information sur l'occurrence du mot améliore la qualité d'indexation, une fois comparée à la fréquence inverse du document. Cette expérience indique également un lien entre les mots « porteurs de sens » et les mots qui ont tendances à se regrouper. Malheureusement les auteurs ne précisent pas comment employer ces mesures pour caractériser les thèmes dans un document (catégoriser les textes), ou quels sont les mots associés à ces thèmes.

Brewster dans [1] définit un canal comme étant le reflet de la pertinence des mots du texte par rapport à un thème donné. Ils appliquent par la suite une transformée en ondelettes de Haar sur la collection des signaux résultante (ensemble de canaux) afin de construire un signal appelé signal d'énergie composée qui permet d'identifier les ruptures thématiques à travers le document (niveau fréquence), ces ruptures peuvent être analysés à différents degrés de détails. Cette modélisation fut la base de construction d'un outil de visualisation des ruptures thématiques dans des documents électroniques non structurés [3]. Précisons qu'aucun résultat expérimental concernant le calcul de l'énergie composée, n'a fait l'objet d'une publication à l'heure actuelle.

L'implémentation mathématique de la transformée de Haar est définit par :

$$W_{j,k,m} = {}^{1}\!\!/\!_{2^{k/2}} \left[\sum_{j'=1}^{2(k-1)} Y_{-j'+j,m} - \sum_{j'=1}^{2(k-1)} Y_{j'-1+j,m} \right]$$

Où W est l'énergie de canal, Y représente le vecteur d'associations, M représente un Canal (thème), K le niveau de multi résolution (de 1 jusqu'à K) et J correspond à l'index du document (l'ensemble des mots porteurs de sens).

Le calcul de l'énergie composée revient à généraliser le calcul de l'énergie de canal pour touts les thèmes (m), pour une position fixe dans l'index J et pour un niveau de multi résolution fixe (k).

3 Utilisation de la méthode des mots associés pour la modélisation du texte en signal

Un texte est un ensemble de thèmes, dont chacun peut être abordé dans une ou plusieurs parties et le flot de chaque thème reflète le degré de pertinence des mots du texte par rapport à ce thème dans le texte. Supposons que l'on puisse mesurer la pertinence d'un mot w par rapport à un thème donné T en utilisant une mesure D (w, T). Pour observer le flot de ce thème on trace la courbe représentant le niveau de pertinence des mots du texte par rapport à ce thème pris dans leur ordre chronologique. Il convient donc de définir une méthode permettant d'extraire et de représenter les thématiques pertinentes dans un texte. Le processus de sélection thématique définit par Miller [3] étant décrit partiellement, il s'avère difficile à implémenter dans la pratique. Dans cet article nous proposons d'utiliser la méthode des mots associés, basée sur la cooccurrence des mots clés et qui permet d'identifier les expressions les plus fortement associées, conduisant à mots candidats pertinents par rapport à différentes thématiques dans les textes à analyser.

3.1 Méthode des mots associés

La méthode des mots associés est une technique développée au début des années 80 par le CSI[1] de l'école de Mines, et le CDST[2] du CNRS [3][19,20]. Cette méthode est née des problèmes posés par l'application des méthodes statistiques usuelles aux données utilisées par la sociologie des sciences. Elle propose d'identifier les mots les plus fortement associés.
L'approche des mots associés concerne l'analyse de co-occurrences de termes indexés et permet l'expression des convergences d'intérêts entre acteurs et actants. L'association des deux mots clés se mesure en fonction de leur simultanéité d'apparitions dans les textes qu'ils indexent. De récents travaux [21] illustrent la pertinence de cette méthode.

3.2 Expérimentation

Cette expérience comporte deux étapes :

- Une analyse statistique du corpus en utilisant la méthode des mots associés, dont l'objectif est de déterminer les thèmes principaux du corpus.
- Un traitement mathématique qui permettra de construire le signal correspondant a chaque flot thématique dans le corpus.

3.2.1 Etape 1

Dans cette expérimentation, nous allons délibérément ignorer la question de choix des thèmes. La technique choisit s'appliquera de façon identique sur tous les thèmes. Un thème est un sujet ou une catégorie prédéfinie (le lecteur pourra se référer à titre d'exemple aux catégories choisies par Yahoo ou bien aux classes utilisées dans la base de données Reuters [23]).

[1] CSI : Centre de Sociologie de l'Innovation
[2] CDST : Centre de Documentation Scientifique et Technique
[3] CNRS : Centre National de la Recherche Scientifique

Cet exemple utilise la base des archives du parlement européen. La classification utilisée par cette base propose un ensemble de catégorie prédéterminé sous lesquelles les documents sont classés manuellement. Chaque document est classé sous une ou plusieurs des onze catégories de la base.

Catégories
Citoyenneté européenne
Culture et education
Emploi et sociale
Economie et monnaie
Société de l'information

Les documents sont au format texte. Notre corpus contient 15 documents. Ces documents apparaissent dans l'ordre chronologique où ils ont été introduits dans le projet. Leur taille varie entre 2 et 10 pages.

Chaque document est analysé sous forme d'une suite de mots. Les caractères de fin de ligne présents dans un document divisent le document en autant d'extraits (segments). La notion d'extrait servira de base pour la recherche de cooccurrences et la formation de clusters. Un document peut contenir des commentaires (par défaut, de la forme {{texte entre commentaires}}). Le contenu de ces commentaires est supprimé à l'analyse du document, et remplacé par un groupe vide {{E}}.

Considérons l'ensemble des articles de notre corpus. Les premiers traitements simples consistent à établir la liste des mots utilisés et de calculer leurs fréquences (Index). Nous utiliserons pour cette étape l'extracteur terminologique intégré au progiciel Sampler.

L'extracteur terminologique de Sampler est composé d'un lexique, d'une liste de patrons et de règles de désambiguïsation contextuelles. Le lexique est constitué d'unitermes étiquetés par leur catégorie grammaticale (adjectif, verbe, préposition, article, nom, conjonction, ...).

Les paramètres d'extraction sont fixés comme suit :
Fréquence minimale d'extraction : 1
Nombre de mots par expression : 4

3.2.1.1 Constat

L'index produit par Sampler contient 1800 mots, d'occurrence variant ente 1 et 43. Un examen rapide de l'index permet de situer le thème principal traité dans le corpus. Les premières expressions sont : européen, travailleurs, droit, mobilité, etc…

3.2.2 Etape 2

Le fondement de la méthode des mots associés est la notion de co-occurrences de mots dans l'ensemble des documents. Il s'agit aussi de définir un (ou des) indice(s) pour mesurer l'intensité relative de ces cooccurrences.

De deux mots i et j on dit qu'ils co-occurrent ou qu'ils sont associés s'ils sont utilisés ensemble pour décrire un même document.

Il existe plusieurs réalisations informatiques de cette méthode. La plus ancienne est le logiciel Leximappe [10] développée conjointement par le CDST du CNRS, et le CSI De l'Ecole de Mine de Paris.

A partir de l'index construit précédemment par Sampler, on construit la **matrice de cooccurrences** « Mots index * Mots index ».

$$\begin{pmatrix} 29 & 11 & 7 & \cdots & \cdots \\ 11 & 32 & 4 & & \\ 7 & 4 & 17 & & \\ \vdots & \vdots & \vdots & & \\ \vdots & \vdots & \vdots & & \ddots \end{pmatrix}$$

Tableau 1. Exemple de la matrice de co-occurrences

3.2.2.1 Interprétation

La co-occurrence ne permet pas à elle seule de mesurer la force et la pertinence des associations entre les mots, car elle avantage les mots qui co-occurrent un grand nombre de fois dans le corpus.

L'emploi d'un indice statistique permet de normaliser cette mesure dont il existe plusieurs variantes. Le coefficient utilisé dans la méthode des mots associés est l'indice d'équivalence E_{ij} [20] défini par :

$$E_{ij} = \frac{C_{ij}^2}{C_i * C_j}$$

Où :

- C_{ij} le nombre de co-occurrences des mots i et j.
- C_i le nombre d'occurrences du terme i.
- C_j le nombre d'occurrences du mot j.

On va ainsi mesurer l'intensité de l'association existante entre les mots i et j. Si E_{ij} =1, cela signifie que la présence d'un terme entraîne la présence de l'autre, si E_{ij} =0, cela signifie que la présence d'un terme exclu la présence de l'autre dans le corpus.

1	0,13	0,10	...	...
0,13	1	0,02		
0,10	0,10	1		
⋮	⋮	⋮		
⋮	⋮	⋮		⋱

Tableau 2. Extrait de la matrice calculée à partir de l'indice d'équivalence

Cette méthode va générer un nombre important de liens entre les couples de mots. Partant de la matrice normalisée, on classe par ordre décroissant le poids des couples de mots. On obtient le résultat suivant :

Mot 1	Mot 2	Poids de la relation
européen	Pays	**0,13**
européen	Travail	**0,10**
Travail	**Pays**	**0,02**

Tableau 3. Classement par ordre décroissant de poids de relation entre mots

Partant du Tableau 3, nous allons constituer un tableau des mots les plus centraux ayant les poids de relation les plus importants.

Mot 1	Pois de la relation
Européen	**0,13+0,10=0,23**
Pays	**0,13+0,02=0,15**
Travail	**0,10+0,02=0,12**

Tableau 4. Liste des mots les plus centraux ayant une forte pertinence (Ordonnée par les valeurs décroissante de pertinence)

3.2.3 Etape 3 : Construction du signal du flot thématique

Le flot thématique reflet le degré de pertinence d'un thème aux divers points dans un document. Une représentation idéale du flot thématique devrait permettre la présence de plusieurs thèmes simultanément à un point donné dans un texte, cette flexibilité peut être réalisée en représentant indépendamment les thèmes les uns les autres.

L'algorithme de construction parcourt le texte et pour chaque position du mot w, il attribue pour la position correspondante dans le signal, la valeur reflétant la pertinence de ce mot par rapport au thème T choisi. A partir de la liste des mots correspondant à un thème donnée (cluster) ordonnés par leurs poids de similarité, l'algorithme procède comme suit : Si le mot w appartient à la liste correspondante au thème T, il lui sera assigné une valeur correspondante à son degré de similarité par rapport à ce thème (valeur du lien interne). Si w n'appartient pas à la liste de T, il lui sera assigné la valeur du lien externe avec le thème T', sinon il sera rejeté. Le flot de chaque thème peut être présenté séparément de la même manière, en supposant que plusieurs thèmes soient présents simultanément dans un texte.

La représentation du flot thématique du thème « politique de l'emploi » (figure. 3) permet d'observer des variations des valeurs de pertinence plus importantes que celles observées dans le signal du flot de « citoyenneté européenne » (figure. 4). Ces variations importantes indiquent une présence relativement plus faible des mots de ce dernier thème, donc une présence relativement faible du thème dans ce segment.

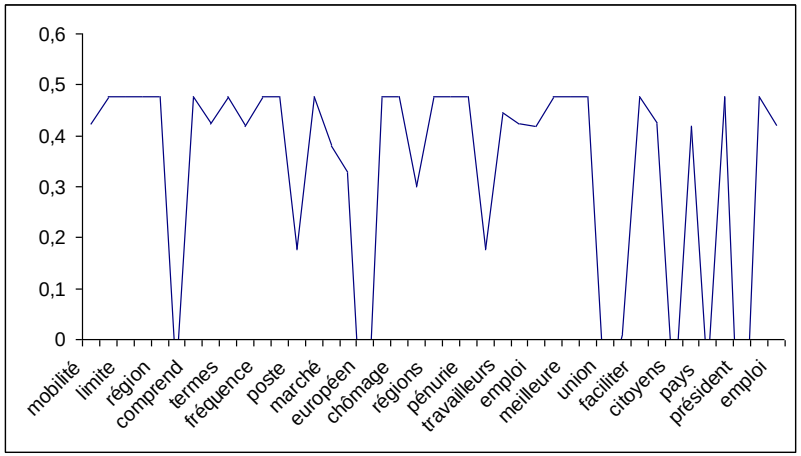

Figure 3. Flot thématique du thème
« Politique de l'emploi »

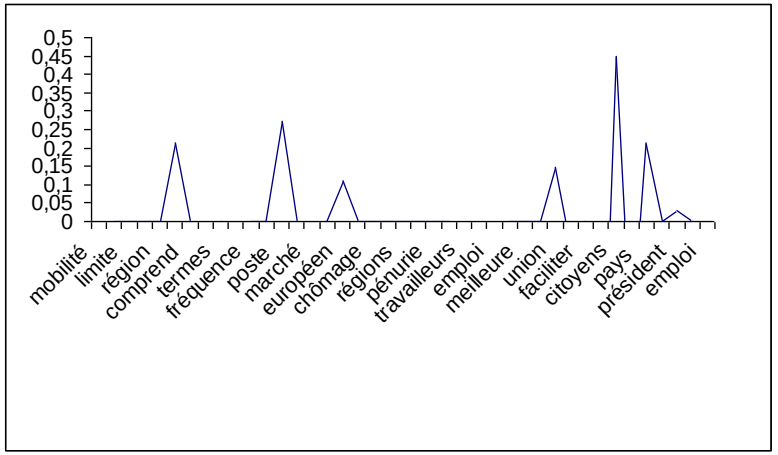

Figure 4. Flot thématique du thème
« Citoyenneté européenne »

4 Conclusion

L'approche considérant les termes indépendamment de leur apparition chronologique dans les documents « bag of Words » suppose que les mots utilisés suffisent pour donner le sens du contenu général du document. Elle suppose également que la probabilité d'utiliser un mot est indépendante des autres mots et de sa position dans ce document. Selon cette approche, le texte est une large entité sans aucune structure. C'est l'approche choisie et mise en œuvre actuellement dans la plupart des SRI. Ceci est sans aucun doute lié à la simplicité d'implémentation de cette méthode. Beaucoup de systèmes augmentent leurs performances en couplant cette approche à une analyse morpho-syntaxique régler partiellement les problèmes de synonymie. Certains systèmes récents améliorent également ce dispositif en s'appuyant sur la sémantique distributionnelle.

Dans cet article, nous avons présenté une méthode de construction de signaux pour représenter le flot thématique d'un document. Nous avons tout d'abord exposé une méthode de sélection de thèmes pertinents, avant de présenter un algorithme de construction du signal du flot basé sur des mots candidats. La qualité du signal construit repose sur la qualité des mesures de sélection et de calcul de pertinence utilisées dans ces travaux. L'originalité de cette recherche repose pour partie sur l'utilisation de l'indice d'équivalente dans cette optique.

Dans un prochain article, nous proposerons la généralisation de la modélisation du contenu textuelle en signal pour des segments de tailles variables. Nous étudierons par la suite l'effet de certaines méthodes de lissage, en particulier, les transformée en Ondelettes dans le but de réduire le bruit et clarifier le signal du flot. Nous présenterons également les applications possibles de cette modélisation à des fins de recherche et la comparaison documentaire.

5 Bibliographie

[1] BREWSTER ET AL, *Information Retrieval System Utilizing Wavelet Transform*, Battelle Memorial Institute, Richland, 2000.

[2] Reem Al-HALIMI, *Mining Topic signals from Text*, Canada, 2003.

[3] MILLER N ET AL, *Topic Islands- A Wavelet Based Text Visualization System*, Pacific Northwest National Laboratory (http://www.pnl.gov/), IEEE, 1998.

[4] BOOKSTEIN ET AL, *Clumping Properties content-Bearing Words*, Journal of The American Society for Information Science, 1998.

[5] DAS-GUPTA P, *Boolean interpretation of conjunctions for document retrieval*, Journal of the American Society for Information Science, 1987.

[6] SAVOY J, *Ranking schemes in hybrid Boolean System: A new approach*, Journal of the American Society for Information Science (http://www.asis.org/Publications/JASIS/jasis.html), 1997.

[7] SALTON G, *The SMART Retrieval System – Experiments in automatic document processing*, Prentice- Hall, 1971.

[8] RAJMAN M, BONNET A., *Corpora-base linguistics: new tools for natural language processing*, 1st Annual conference of the Association for Global Strategic Information, Allemagne, 1992.

[9] DEERWESTER S ET AL, *Indexing by latent semantic indexing*, Journal of the American Society for Information Science, 1990.

[10] MICHELET B, *L'analyse des associations*, 1988.

[11] JOUVE O, *Les outils d'analyse et de filtrage d'informations : l'exemple du projet SAMPLER*, Pole universitaire Léonard de Vinci, Paris.

[12] *Nuclear Proliferation News*, ACRONYM Consortium, Sean Howard, Issue N° 33, 27 September 1995.

[13] CHRISTOPHER D, *Foundations of Statistical Natural Language Processing*, the MIT Press, Cambridge, Massachusetts, 1999.

[14] BOOKSTIEN A, Fuzzy requests: *An approach to weighted Boolean searches*, JASIS, 1980.

[15] SALTON G et Fox E.A *Extended Boolean Information Retrieval*, Communication ACM, 1983.

[16] GAUSSIER E ET AL, *Recherche d'information et traitement automatique des langues : expériences sur le français et en recherche d'information multilingue*, T.A.L, vol 41, N°2.

[17] MALLAT S, *Une exploration des signaux en ondelettes*, Edition de l'école polytechnique, novembre 2000.

[18] *BESANÇON R, Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles de textes*, thèse de doctorat, École Polytechnique Fédérale de Lausanne.2001.

[19] M. CALLON, J.P. COURTIAL, W. TURNER, S. BAUIN, *an introduction to co-word analysis*, Social Science Information n°22, 1983.

[20] M. CALLON, F. BASTIDE, S. BAUIN, J. P. COURTIAL, W.TURNER, *Les mécanismes d'intéressement dans les textes scientifiques*, Cahiers S.T.S, N°4, 1984.

[21] B DELECROIX, R EPPSTEIN, *Co-word analysis for the non-scientific information example of Reuters Business Briefings*, Data Science Journal Vol 3, 2004.

[22] RADECK T, *Fuzzy set theoretical approach to document retrieval*, Information Processing and Management Vol 15, 1997.

[23] *The Reuters- test collection* http://www.research.att.com/ lewis/reuters21578.html.