Un système de calcul des thèmes de l'actualité à partir des sites de presse de l'internet

JACQUES VERGNE
GREYC - UMR 6072
campus II - BP 5186
Université de Caen
14000 Caen, FRANCE
mail :Jacques.Vergne@info.unicaen.fr
tél. : 02 31 56 73 36 fax : 02 31 56 73 30

Résumé

Dans cet article, nous présentons un système de constitution de revue de presse à partir des sites de presse présents sur l'internet[1]. Il s'agit de répondre à des questions telles que : "de qui, de quoi est-il question aujourd'hui dans la presse de tel espace géographique ou linguistique ?". L'utilisateur, qu'il soit un journaliste qui prépare sa revue de presse, ou simplement une personne intéressée par l'actualité, définit en entrée l'espace de recherche qui l'intéresse. Ce système inverse la problématique des moteurs de recherche : au lieu de rechercher des documents à partir de mots-clés qui représentent des thèmes, il s'agit de produire en sortie les thèmes principaux de l'actualité, et de donner accès aux articles concernés par ces thèmes. Les thèmes d'actualité sont capturés en relevant les termes récurrents dans les "textes" d'hyperliens des "Unes" des sites de presse. Le système calcule un graphe de termes dans lequel les nœuds sont les termes et les arcs sont les relations entre termes, relations définies par la co-occurrence de deux termes dans un "texte" d'hyperlien. L'interface exploite ce graphe en permettant à l'utilisateur de naviguer parmi les termes et d'avoir accès aux articles contenant ces termes[2].

Mots-clés : hypertextes, web, internet, documents électroniques, web mining, recherche d'informations, veille stratégique, fouille de textes.

Abstract

In this paper, we present a system for building a news review, from news sites on the web. We want to be able to answer questions as : "who, what are papers speaking about today in the news of a given geographic or linguistic search space". The user, a journalist preparing his news review, or somebody interested in news, defines as input the search space he is interested in. This system reverses the issues of search engines : in spite of searching documents from key-words which represents topics, we want to produce as output the main topics of the news, and to give access to related papers. News topics are captured while computing recurrent terms in hyperlinks texts of front-pages of news sites. The system computes a graph in which nodes are terms and arcs are links between terms; a link is defined as a co-occurrence of two terms in a same link text. The interface is based on this graph as the user can browse through the terms and have access to papers containing these terms.

[1] Une démonstration est accessible sur :
http://www.info.unicaen.fr/~jvergne/demoRevueDePresse/index.html

[2] Le système présenté a des analogies avec celui de Google News (http://news.google.fr), mais Google News n'a pas encore publié sur son processus de traitement.

Key-words : hypertexts, web, internet, electronic documents, web mining, information retrieval, strategic watching, text mining.

1 Introduction

Le système que nous présentons comporte en donnée une liste aussi large que possible des URL des sites de presse du monde entier, qui constituent les points d'entrée possibles. L'utilisateur définit en entrée un espace de recherche géographique et/ou linguistique sous la forme d'un sous-ensemble de cette liste. Sa requête implicite est : "de qui, de quoi est-il question aujourd'hui dans la presse de cet espace ?". Le système fournit en sortie un graphe de termes valués, reliés par des relations valuées. Un terme est valué par trois grandeurs : le nombre de termes auxquels il est relié, le nombre de sites sur lesquels il a été trouvé, et le nombre de textes de liens dans lesquels il a été trouvé (ce qui correspond au nombre d'articles concernés par le terme). La relation entre deux termes est définie par leur co-occurrence dans un même "texte" d'hyperlien. Une relation est valuée par le nombre de textes de liens où les deux termes sont présents. Ces valeurs attribuées aux termes et à leurs relations permettent de les classer pour les présenter à l'utilisateur par ordre de présence décroissante dans l'actualité du jour dans l'espace de recherche défini en entrée. Elles permettent aussi des traitements de graphe particuliers. L'utilisateur prend connaissance des résultats en naviguant dans le graphe : il choisit un terme, puis des termes liés, et a accès à tout moment aux documents concernés par les termes et leurs liens.

Nous présentons d'abord les principes de fonctionnement, puis ensuite le processus général de calcul et ses étapes successives.

2 Principes de fonctionnement

Pour chaque site, **un seul** document est téléchargé : le document du point d'entrée, c'est-à-dire la "Une" du site de presse. De ce document, sont extraits les hyperliens : les URL et le code source des "textes" de liens. On observe que ces codes source de "texte" de liens sont composés de titres ou de résumés d'articles (avec leur mise en forme), et d'URL vers des images ou des photographies. Dans une même Une, une même URL peut apparaître plusieurs fois. Les URL de photographies permettent de les montrer à l'utilisateur en sortie.

C'est dans les "textes" de liens (leur code source débalisé) que sont extraits les termes. Les URL des articles ne serviront qu'en sortie, pour donner accès à un article, si l'utilisateur le décide. Le système ne se sert pas des articles eux-mêmes. Cette économie de traitement s'appuie sur le fait que la rédaction d'un texte de lien est un choix éditorial des journalistes des sites de presse.

Un point délicat est la **méthode d'extraction des termes** à partir du corpus des textes de liens débalisés. La tâche est relativement simple : il s'agit de trouver les motifs répétés, dans un corpus thématiquement varié, relativement petit (environ 90 à 160 Ko, 17000 à 25 000 mots), tout en repérant les mots grammaticaux (motifs de hautes fréquences), pour éviter d'en faire des termes. Mais nous avons des contraintes particulières sur la méthode : elle doit être robuste et indépendante des langues, ce qui est une nécessité pour un logiciel de traitement d'informations sur l'internet, caractérisé par la multiplicité des langues; la méthode ne doit pas utiliser de ressources propres à une langue, car il n'est pas question de faire un travail de préparation de ressources linguistiques à chaque nouvelle langue traitée.

Nous n'avons donc pas besoin de méthode lourde avec analyse syntaxique et mises en relation, telle que celle de Didier Bourigault qui utilise l'analyseur syntaxique en dépendance SYNTEX pour construire une ontologie à partir d'un corpus vaste et très cohérent thématiquement (voir [Bourigault, 2000] et [Bourigault, 2002]).

Les méthodes d'André Salem (voir [Salem, 1987]) et d'Helena Ahonen (voir [Ahonen-Myka, 1999] et [Ahonen-Myka, 2002]) recherchent les motifs répétés en utilisant des algorithmes extrapolés de l'algorithme glouton (recherche des n-grammes à partir des n-1-grammes), mais ces méthodes utilisent en entrée les mots grammaticaux de la langue traitée pour éviter de les prendre comme termes (stopword-list).

Plusieurs méthodes ont été explorées, en nous imposant la contrainte de trouver une méthode n'utilisant pas de ressources linguistiques, pour rester robuste et indépendant des langues : la recherche des motifs répétés par l'algorithme glouton a été expérimentée, d'abord avec puis sans l'utilisation des majuscules[3], avec exclusion des mots grammaticaux par leur fréquence (test de Zipf). Puis une méthode tout à fait originale fondée sur la périodicité des longueurs de mots a été mise au point. Elle permet de calculer si un mot est grammatical ou lexical sans autre ressource que le corpus traité lui-même (une méthode dite "endogène", en reprenant le terme de Didier Bourigault).

Un fois les termes calculés, on leur associe la liste des sites où on les a trouvé, et on ne garde que les termes trouvés sur au moins deux sites, donc dans des textes de liens vers deux articles différents de deux sites différents. C'est cette opération d'intersection qui permet d'exclure les termes particuliers à un site unique (menus, publicités, etc.).

On associe ensuite à chaque terme restant sa liste de liens vers les articles.

Les **relations entre termes** peuvent ensuite être calculées : il existe une relation entre deux termes s'ils sont co-occurrents dans un même texte de lien, ou, ce qui est équivalent, si l'intersection des deux listes d'articles n'est pas vide. Dans cet étape, il s'agit simplement de calculer les intersections de listes de liens des termes deux à deux.

Étant donné que plusieurs centaines de termes sont extraits, il faut, pour faciliter l'accès à l'utilisateur, les lui présenter par **groupes de termes fortement reliés** : les groupes calculés par le système à partir de propriétés du graphe, sont interprétés par l'utilisateur comme des groupes de termes thématiquement reliés (cf. le "grouping process" de Google News[4]). Une idée simple et classique est de calculer les composantes connexes du graphe de termes (c'est-à-dire ses sous-graphes non connexes). Cette solution a été expérimentée, puis abandonnée, car le graphe est composé d'une très grosse composante (presque tous les thèmes de l'actualité sont liés), et de beaucoup de petites composantes (un thème a donné 2 termes reliés ou même un seul terme). Donc le problème à résoudre est de segmenter la composante connexe principale en ensembles de nœuds fortement reliés. Là encore, une solution classique se présente : la recherche des cliques maximales[5]. Cette solution a aussi été expérimentée, et aussi abandonnée : l'algorithme glouton (recherche des cliques de n nœuds à partir des cliques de n-1 nœuds) donne un résultat combinatoire qui pose le problème supplémentaire de choisir entre des cliques équivalentes. Mais cette méthode a surtout un intérêt sur les graphes non valués. Or nous sommes en présence de graphes valués, d'où la recherche d'une méthode de groupage des termes fondée sur les valeurs des liens et sur les valeurs des nœuds : les groupes sont constitués à partir des nombres de co-occurrences des termes et de leurs nombres de sites.

Il s'agit enfin de présenter à l'utilisateur une **interface d'accès au graphe des termes et aux articles**. Une solution envisageable est un graphe cliquable, sous-graphe du graphe des termes, à la manière de Kartoo (`www.kartoo.com`). Nous avons choisi une solution plus simple à mettre en œuvre (Kartoo utilise la technologie flash) : le graphe des termes est transposé dans un graphe de documents html (un document html par nœud-terme) reliés par des hyperliens dans les deux sens (2 hyperliens réciproques instancient un arc non orienté entre 2 termes). L'utilisateur peut ainsi naviguer dans le graphe des termes.

Ces principes de fonctionnement ont permis de construire un système léger, robuste, sans ressources linguistiques, indépendant des langues, qui utilise des propriétés très générales des langues.

3 Le processus général et ses étapes

Voici les étapes du processus général des traitements :

- phase préparatoire manuelle : collecter les URL des sites de presse

[3] Pour pouvoir traiter des langues sans majuscules, comme l'arabe par exemple.

[4] Voir sur `http://news.google.com/help/about_news_search.html` : "an automated grouping process for Google News that pulls together related headlines and photos from thousands of sources worldwide".

[5] Une clique (ou graphe complet) est un ensemble de nœuds où tout nœud est relié à tous les autres.

- traitements sur l'ensemble des sites :
  - télécharger et analyser **la Une de tous les sites**
  - relever **les termes** dans les textes de liens
  - pour chaque terme, calculer sa liste de sites et sa liste d'articles

- entrer l'espace de recherche de l'utilisateur

- traitements sur les sites choisis par l'utilisateur :
  - calculer **les relations entre termes**
  - regrouper les termes fortement reliés
  - sortir les résultats = calculer l'interface

3.1 Collecter les URL des sites de presse (phase préparatoire)

Les URL des sites de presse ont été collectées sur les sites suivants :

- le Courrier International propose une liste commentée d'environ 800 sites de presse du monde entier présentée par continent (sauf la France) :
  `http://www.courrierinternational.com/kiosk/`
- Google News donne une centaine de ses sources sur les 4000 revendiquées :
  `http://news.google.fr/news/`
- le site de NewsLink donne aussi de nombreux sites de presse :
  `http://newslink.org/`

Notre système est testé quotidiennement sur une quarantaine de sites : 22 de la presse française nationale et régionale, 17 de la presse européenne (Suisse, Belgique, Allemagne, Italie, Espagne, UK, Irlande), et 4 sites de presse nord-américaine (espace 1 dans la suite de l'article). Des tests de vitesse d'exécution sont aussi fait sur les 100 sites de Google News, environ la moitié sont des sites nord-américains (espace 2 dans la suite de l'article).

Cette phase s'apparente au "sourcing" des sociétés de veille technologique.

3.2 Traitements sur l'ensemble des sites

3.2.1 Télécharger et analyser la Une de tous les sites

Le téléchargement des Unes est fait périodiquement (une fois par jour, par exemple), pour l'ensemble des sites de la base de sites. Le téléchargement des Unes et le relevé des termes dans les textes de liens sont dissociés de l'interrogation par l'utilisateur. Les phases suivantes, à commencer par le calcul des relations entre termes, sont dépendantes de l'espace de recherche défini par l'utilisateur.

Après téléchargement, le code source de la page est analysé. Pour chaque lien, les URL et le code source des textes de liens sont extraits[6]. Les URL relatives sont converties en URL absolues. Comme une même URL peut apparaître plusieurs fois, les textes de liens sont concaténés, et les URL d'image sont aussi converties en URL absolues. Après ces conversions, les URL et les codes sources des liens sont mémorisés dans une base de données.

3.2.2 Relever les termes dans les textes de liens

Cette étape consiste à extraire des termes d'un corpus : le corpus des textes de liens concaténés, obtenus par débalisage des codes source des liens, entre les balises `<a>` et `</a>`. Pour fixer les idées, ce corpus est de l'ordre de 90 Ko pour les 43 sites de l'espace de recherche 1.

Une première question se pose au sujet de **l'espace de constitution du corpus**, c'est-à-dire l'espace dans lequel on va rechercher des motifs répétés. Plusieurs solutions sont possibles : un corpus pour tous les sites, un corpus par langue (si on décidait de se servir de la langue de chaque site), ou un corpus par site. Nous avons choisi un corpus pour tous les sites, ce qui permet dès cette étape, de ne considérer que les termes répétés (au moins 2 occurrences) dans l'ensemble du corpus, qui pourront

[6] Il est relativement fréquent que les balises `<a>` ne soient pas fermées par une balise `</a>`. Dans ce cas, on utilise une fermeture implicite par la prochaine balise `</p>` ou `</td>`. De plus, pour pallier les liens quasi-vides (tels que "*Lire*"), on concatène au texte de lien le texte du `<td>` englobant le lien.

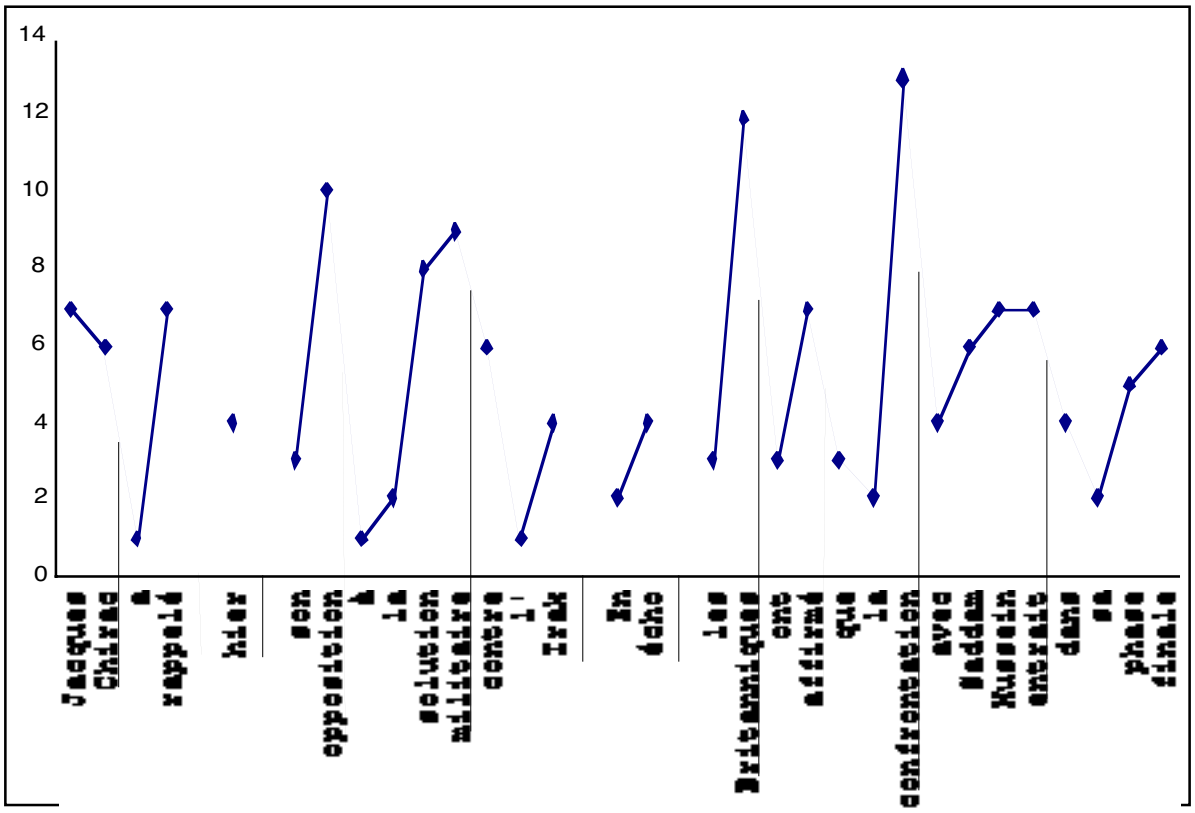
éventuellement se trouver sur deux sites différents; alors que le calcul des termes par site oblige à garder temporairement les termes hapax, car deux termes hapax dans deux sites différents peuvent être identiques, et donc répétés sur l'ensemble des sites; or garder les termes hapax temporairement alourdit inutilement les calculs.

Une deuxième question est celle de la **méthode d'extraction des termes** à partir d'un corpus.

Dans un premier temps, nous avons expérimenté l'algorithme glouton, associé avec un test de Zipf pour éliminer les mots grammaticaux des termes, sur un simple critère de fréquence. Les résultats étaient corrects, mais le départage entre les mots grammaticaux et les mots lexicaux très fréquents était difficile (tel que "guerre" aujourd'hui).

D'où une autre manière de poser le problème : comment distinguer les mots lexicaux et les mots grammaticaux à partir du corpus ? Une direction de travail est d'utiliser à la fois les formes et leurs positions (une constante du Groupe Syntaxe du GREYC, dont Hervé Déjean [Déjean, 1998], Nadine Lucas [Lucas, 2001]), alors que le test de Zipf n'utilise que les fréquences des formes, sans exploiter leurs positions relatives. Or une observation fondamentale de Zipf est que les mots grammaticaux sont fréquents et courts et que les mots lexicaux sont plus rares et plus longs (ce qui est d'usage fréquent est court : c'est la loi de l'économie d'effort dans l'usage d'un code, caractérisée par Zipf, et visible aussi dans les langages de programmation). D'où l'idée d'utiliser non plus les différences de fréquence, mais les différences de longueur. Si on associe les longueurs des mots (un type de forme) avec les positions relatives des mots, on est tout naturellement conduit à s'intéresser aux variations des longueurs au fil du texte. Voici en figure 1 un exemple de graphe de la fonction :

Figure 1 : Graphe de la fonction :
nombre de lettres d'un mot = f(position du mot au fil du texte)
(exemple extrait de Ouest-France du 22 février 2003)



Le texte est segmenté sur les ponctuations, d'où des segments physiques que nous nommons "virgulots". Dans chaque virgulot, on observe quelques périodes d'**un signal périodique**, avec des alternances court-long avec le plus souvent un début de virgulot court; on observe aussi que la segmentation en chunks est manifestée par un début court, soit en début de virgulot, soit après un mot

long. Sur ces propriétés, nous avons bâti un algorithme de chunking approché et de catégorisation des mots en trois catégories : les mots courts, les mots longs, et les mots indéterminés. Cet algorithme consiste à couper aux endroits où la pente est négative, ce qui revient à garder les parties monotones croissantes (motif court*-long+), qui correspondent à un chunk, c'est-à-dire à un groupe accentuel : on obtient un segment caractérisé par sa prosodie, propriété linguistique indépendante des langues (voir la présentation prosodique du chunk chez Abney, au début de [Abney, 1991]).

Puis, à l'intérieur de chaque chunk, l'algorithme consiste à catégoriser chaque mot : court ou long, suivant sa longueur et sa position. Puis les déductions locales aux chunks sont consolidées par une étude globale sur tout le corpus : ne sont retenus que les mots de nombre d'occurrences supérieur à 1 (les seuls qui pourront devenir des termes) et de catégorie stable sur le corpus (la "stabilité" se caractérisant par une majorité pour une catégorie). À partir de cette catégorisation, deux motifs de termes sont possibles (avec le langage des expressions régulières, c pour **c**ourt, L pour **L**ong) :

`c*L+` *(la) [violence] routière*

`c*L+cL+` *(l)'[intervention] (de) [Jean-Pierre Raffarin]*

Le terme (`L+` ou `L+cL+`) est mémorisé avec tous ses contextes (leurs chunks), pour pouvoir les fournir à l'utilisateur en sortie (par coloriage).

Enfin, certains termes sont fusionnés sur la ressemblance de leurs graphies :

ONU ≡ Onu, Côte d'Ivoire ≡ Côte-d'Ivoire

Cette méthode a de nombreux avantages : pas de ressources externes, pas de diagnostic de langue, exploitation de propriétés linguistiques indépendantes des langues, robustesse, insensibilité aux déséquilibres entre langues dans le corpus, étant donné qu'on n'utilise pas de test de Zipf [7].

3.2.3 Pour chaque terme, calcul de sa liste de sites et de sa liste d'articles

Pour chaque terme, on calcule sa liste de sites en vérifiant sa présence dans le corpus de chaque site. À ce stade, les termes présents sur un seul site sont éliminés.

Pour chaque terme, pour chaque site où est ce terme, on calcule enfin sa liste d'articles en vérifiant sa présence dans chaque texte de liens de ce site.

3.3 Entrer l'espace de recherche de l'utilisateur

L'utilisateur définit interactivement par un formulaire son espace de recherche, c'est-à-dire un sous-ensemble de l'ensemble des sites de presse entrés dans la phase préparatoire. Les critères sont linguistiques et/ou géographiques. Chaque site doit être complété de ces deux informations, qui ne servent qu'à cette étape.

3.4 Traitements sur les sites choisis par l'utilisateur

Au moment de l'entrée de l'espace de recherche de l'utilisateur, la liste d'articles de chaque terme est filtrée pour ne retenir que les sites choisis.

3.4.1 Calculer les relations entre termes

À partir de cette étape, les calculs sont dépendants de l'espace de recherche défini par l'utilisateur.

Il existe une relation entre deux termes si l'intersection des deux listes d'articles n'est pas vide. La valeur d'une relation est définie par le cardinal de l'intersection.

Dans cette étape, il s'agit simplement de calculer les intersections de listes d'articles deux à deux, ce qui est possible à l'aide de requêtes sur la base de données. On obtient le graphe (non orienté) des termes.

On peut donner ici quelques informations quantitatives sur les graphes de termes obtenus : 400 à 700 termes-nœuds, 1600 à 2500 relations-arcs, d'où des densités (rapport entre le nombre d'arcs réels et le nombre d'arcs possibles) de l'ordre de 1%.

[7] Voir aussi dans [Vergne, 2003], une extension de cette méthode, avec utilisation des différences de longueur **et** d'effectif dans le corpus des "textes" de liens.

3.4.2 Regrouper les termes fortement reliés

Pour regrouper les termes, on parcourt la liste des couples de termes liés par valeurs décroissantes des relations (jusqu'à une valeur 2 : on néglige à ce stade les relations trop faibles), et on place les 2 termes a et b dans un groupe avec les règles suivantes :

- si un groupe contient a ou b ou les 2, placer a et b dans ce groupe
- si aucun groupe ne contient a ou b, créer un nouveau groupe et y mettre a et b
- si a et b sont déjà dans 2 groupes différents, mémoriser ce couple.

Les couples ainsi mémorisés sont des relations passerelles entre groupes, à partir desquelles on fusionne quelques groupes fortement liés, en fusionnant d'abord des groupes de tailles très différentes (un petit rejoint un gros). Le problème est de définir un critère d'arrêt, pour éviter de remettre trop de termes ensemble. Le critère actuel est un seuil du rapport de taille des groupes, la taille d'un groupe étant évaluée par la somme des valeurs des relations des couples qui le constituent. Le fonctionnement actuel est correct, mais demande encore à être travaillé.

3.4.3 Sortir les résultats : interface

L'interface permet à l'utilisateur de naviguer dans le graphe des termes et d'accéder aux articles. Le système produit en sortie un document html par terme, avec 2 hyperliens réciproques par relation entre termes.

L'interface est constituée de 2 sous-fenêtres :

- la première propose un choix entre les termes, présentés par groupe, par ordre décroissant d'importance (somme des valeurs des relations), et dans chaque groupe, les termes par nombre de sites décroissants; l'utilisateur choisit un terme en cliquant dessus, et le terme apparaît dans la deuxième sous-fenêtre;
- la deuxième sous-fenêtre permet de naviguer dans le graphe à partir du terme choisi : on y voit ce terme, avec ses termes liés présentés par valeur décroissante de la relation; pour chaque terme lié, sont présentés les articles dans lesquels les deux termes sont co-occurrents; pour chaque article, l'utilisateur voit le lien sur l'article, le texte du lien avec son éventuelle photographie associée et sa mise en forme originale (le terme choisi et ses termes liés sont coloriés pour faciliter leur repérage en lecture rapide); l'utilisateur peut choisir de cliquer sur un terme pour continuer sa navigation dans le graphe des termes, ou il clique sur un lien d'article, ce qui fait apparaître l'article dans une nouvelle fenêtre.

Quelques résultats intermédiaires des calculs et une copie d'écran de l'interface sont donnés en annexe à la fin de l'article.

4 Conclusion

La collecte d'informations sur l'internet, la synthèse des informations collectées, et leur mise à la disposition des utilisateurs, sont des tâches très intéressantes car, tout en étant un enjeu opératoire et social, elles posent aussi de nouveaux problèmes de traitement du matériau linguistique : sur la toile, les langues sont multiples, le lexique est ouvert; la variété et l'immensité du matériau, rendant illusoire l'accumulation de ressources linguistiques propres à des langues particulières, nous conduisent à l'exploitation de propriétés linguistiques de plus en plus générales, et de plus en plus abstraites.

Dans le cas particulier du système présenté dans cet article, nous nous sommes servi (implicitement) de propriétés du groupe accentuel pour ne retenir que des mots lexicaux dans les termes; nous avons beaucoup utilisé la co-occurrence, l'unique ne nous intéresse pas, nous recherchons le multiple : co-occurrence entre 2 sites, puis entre 2 articles; la co-occurrence est une manière d'exprimer la relation : principalement la relation entre termes, qui se traduit dans notre système par un arc du graphe des termes.

Quels sont les acquis ? Le système fonctionne très bien et donne quotidiennement satisfaction à l'auteur comme utilisateur intéressé par l'actualité mondiale. Il permet de mettre à l'épreuve notre nouvelle méthode d'extraction de termes à partir de corpus brut multilingue. Il est frappant de constater combien la recherche multisite est facile, alors que la même tâche en monosite est très difficile (nous nous y sommes attaqué précédemment); c'est une bonne illustration de l'efficacité de la co-occurrence intersite.

Le système présenté nous invite à poursuivre nos recherches dans plusieurs directions : notre méthode d'extraction de termes de corpus brut est à consolider sur des corpus plus importants et sur une plus grande variété de langues (actuellement : français, anglais, allemand, italien, espagnol), et sa robustesse doit être étendue : le choix actuel est de tenter d'annuler le bruit, au prix d'un certain silence, ce qui n'est pas un désavantage dans la tâche actuelle, qui consiste à trouver des termes fréquents, et à ignorer les hapax. De manière plus précise, le chunking sur corpus brut multilingue et sans ressources devra être approfondi. Le calcul des intersections des ensembles de document devra être amélioré car, du point de vue calculatoire, c'est le point faible de la chaîne des traitements : il consiste en un parcours de la demi-matrice carrée des quelque 500 termes (moins la diagonale), soit de l'ordre de 500*500 / 2 = 125 000 calculs d'intersection. Le groupage de termes, qui ne donne pas encore entière satisfaction, est un problème difficile, qui demandera l'exploration de nouvelles solutions. Enfin, l'interface doit évoluer, en interaction avec des utilisateurs.

Références

[Abney, 1991] Abney Steven (1991). "Parsing By Chunks". In: Robert Berwick, Steven Abney and Carol Tenny (eds.), *Principle-Based Parsing*. Kluwer Academic Publishers, Dordrecht.
`http://www.sfs.nphil.uni-tuebingen.de/~abney/Abney_90e.ps.gz`

[Ahonen-Myka, 2002] Ahonen-Myka Helena (2002). Discovery of frequent word sequences in text. *The ESF Exploratory Workshop on Pattern Detection and Discovery in Data Mining*, Imperial College, London, 16-19 September.
`http://www.cs.helsinki.fi/u/hahonen/ahonenmyka_patws02.ps`

[Ahonen-Myka, 1999] Ahonen-Myka Helena (1999). Finding All Frequent Maximal Sequences in Text. *Proceedings of the 16th International Conference on Machine Learning ICML-99 Workshop on Machine Learning in Text Data Analysis*, eds. D. Mladenic and M. Grobelnik, p. 11-17, J. Stefan Institute, Ljubljana .
`http://www.cs.helsinki.fi/u/hahonen/ham_icml99.ps`

[Bourigault, 2002] Bourigault, Didier (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus, *Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2002)*, Nancy, pp. 75-84.
`http://www.univ-tlse2.fr/erss/textes/pagespersos/bourigault/TALN02-Bourigault.doc`

[Bourigault, 2000] Bourigault Didier & Slodzian Monique (2000). Pour une terminologie textuelle, *Terminologies Nouvelles*, n° 19., pp. 29-32.
`http://www.univ-tlse2.fr/erss/textes/pagespersos/bourigault/TN-Bourigault-Slodzian.rtf`

[Déjean, 1998] Déjean Hervé (1998). "Concepts et algorithmes pour la découverte des structures formelles des langues", spécialité : informatique, thèse de l'Université de Caen.

[Lucas, 2001] Lucas Nadine (2001). Étude et modélisation de l'explication dans les textes, *Colloque "L'explication: enjeux cognitifs et communicationnels"*, Paris, 30 novembre - 1er décembre.

[Salem, 1987] Salem André (1987), *Pratique des segments répétés*, Publications de l'INaLF, collection "St.Cloud", Paris, Klincksieck, 333p.

[Vergne, 2000] Vergne Jacques (2000). *Trends in Robust Parsing*, tutoriel du CoLing 2000, Nancy, Sarrebrück.
`http://www.info.unicaen.fr/~jvergne/tutorialColing2000.html`

[Vergne, 2001] Vergne Jacques (2001). Analyse syntaxique automatique de langues : du combinatoire au calculatoire (communication invitée), *Actes de TALN 2001*, 15-29.
`http://www.info.unicaen.fr/~jvergne/TALN2001_JV.ppt.zip`

[Vergne, 2002] Vergne Jacques (2002). Une méthode pour l'analyse descendante et calculatoire de corpus multilingues : application au calcul des relations sujet-verbe, *Actes de TALN 2002*, 63-74.
`http://www.info.unicaen.fr/~jvergne/TALN_2002/TALN2002_JVergne.doc.pdf`

[Vergne, 2003] Vergne Jacques (2003). Un outil d'extraction terminologique endogène et multilingue, *Actes de TALN 2003*, à paraître.

[Zipf, 1949] George Kingsley Zipf (1949), *Human Behavior and the Principle of Least Effort*, Harper, New York, réédition 1966.

Annexe 1 : quelques résultats intermédiaires

Voici les termes les plus fréquents collectés le 26 février 2003 dans chacun des deux espaces définis en 3.1, classés par ordre de nombre de sites décroissants, avec pour chaque terme son nombre de sites et son nombre d'articles :

espace 1 (22 sites de la presse française, 17 de la presse européenne, et 4 sites de presse nord-américaine)			espace 2 (100 sites de Google News)		
Irak (14-29)	jour (7-8)	ONU (6-7)	Iraq (25-48)	Health (13-14)	North (10-11)
France (14-25)	Economie (7-7)	Saddam Hussein (6-7)	2003 (23-28)	Home (12-18)	Update (10-11)
Saddam (13-15)	politique (7-7)	Sports (6-7)	U.S. (20-51)	Terms (12-13)	Council (10-10)
2002 (12-17)	site (7-7)	intermittents (6-7)	Bush (20-30)	Special (11-27)	Services (9-13)
2003 (12-14)	article (6-51)	public (6-7)	Privacy (19-21)	House (11-19)	Information (9-12)
Monde (11-15)	Jacques Chirac (6-11)	semaine (6-7)	Saddam (17-23)	White (11-17)	Media (9-12)
Europe (10-11)	Business (6-9)	Blix (6-6)	Business (16-21)	National (11-16)	Sports (9-12)
Sport (10-11)	News (6-9)	Milan (6-6)	International (15-20)	York (11-15)	America (9-11)
Iraq (9-19)	gouvernement (6-9)	Washington (6-6)	more (14-36)	*this* (11-15)	Travel (9-11)
Bush (9-16)	mort (6-9)	Security (14-23)	Privacy Policy (11-11)	Europe (9-10)	
Chirac (9-16)	Bernard Loiseau (6-8)	dossiers (6-6)	Search (14-22)	South (10-14)	
guerre (9-14)	Blair (6-8)	emploi (6-6)	news (14-20)	UN (10-14)	
mois (9-12)	America (6-7)		Press (13-21)	war (10-14)	
Jacques (7-14)			Help (13-15)	Your (10-13)	
Raffarin (7-12)				Online (10-12)	
Cinéma (7-9)					

On peut observer l'abondance de mots avec initiale majuscule, alors que ces mots ne sont pas favorisés par l'algorithme, ils sont simplement plus fréquents dans ce type de corpus. On observe aussi que l'élimination des mots grammaticaux n'est pas parfaite.

Voici les 22 liens les plus forts entre 2 termes, avec leur nombre de co-occurrences :

espace 1	espace 2
Chirac <—11—> Jacques	Privacy <—11—> Privacy Policy
Chirac <—11—> Jacques Chirac	House <—10—> White
Jacques <—11—> Jacques Chirac	White <—10—> White House
article <—11—> Lire	House <—10—> White House
article <—11—> Lire l'article	Iraq <—8—> U.S.
Lire <—11—> Lire l'article	North <—7—> North Korea
Saddam <—7—> Saddam Hussein	Security <—6—> Council
Irak <—6—> Chirac	Security <—6—> Security Council
Irak <—5—> Saddam	Council <—6—> Security Council
Chirac <—5—> article	Iraq <—5—> UN
Irak <—4—> Jacques	Iraq <—5—> Council
Irak <—4—> Jacques Chirac	U.S. <—5—> Turkey
Jacques <—4—> article	2003 <—5—> TM
article <—4—> Jacques Chirac	U.S. <—5—> U.S. Troops
Raffarin <—4—> Jean-Pierre Raffarin	Saddam <—5—> Hussein
America <—4—> Cup	Saddam <—5—> Saddam Hussein
Côte <—4—> Ivoire	Security <—5—> Homeland Security
America <—4—> America's Cup	Hussein <—5—> Saddam Hussein
Côte <—4—> Côte d'Ivoire	Iraq <—4—> Security
Cup <—4—> America's Cup	Iraq <—4—> House

Ivoire <—4—> Côte d'Ivoire Bové <—4—> José Bové	U.S. <—4—> House Iraq <—4—> Blix

On observe une redondance entre termes simples et termes multiples. En prenant l'exemple de *Jacques Chirac*, cela est dû aux faits qu'il existe d'autres *Jacques* et que *Chirac* existe seul sans le prénom.

Annexe 2 : interface