FILTRAR-S : UN OUTIL DE FILTRAGE SÉMANTIQUE ET DE FOUILLE DE TEXTES POUR LA VEILLE

Nicolas CAMPION (*), Jacques CLOSSON (*), Olivier FERRET (**), Jin SHIN (***), Brigitte GRAU (****), Jean Marc LAZARD (*****), Dhafer LAHBIB (**), Romaric Besançon (**), Jean-Marc FLORET (***), Amar-Djalil MEZAOUR (*****), Xavier TANNIER (****)

nicolas_campion@yahoo.fr, jacques.closson@wanadoo.fr, olivier.ferret@cea.fr, jsn@exakis.com, Brigitte.Grau@limsi.fr, jean-marc.lazard@exalead.com, dhafer.lahbib@gmail.com, romaric.besancon@cea.fr, Jean-Marc.FLORET@exakis.com, mezaour@exalead.com, Xavier.Tannier@limsi.fr

(*) ERAKLE, 42 rue d'Artois, 75008 Paris (France),
(**)CEA, LIST, 18 route du Panorama, BP 6, 92265 Fontenay-aux-Roses (France),
(***)EXAKIS, 40 avenue des Terroirs de France, 75012 Paris (France),
(****) LIMSI, UPR-3251 CNRS-DR4, Bat. 508, BP 133, 91403 Orsay Cedex (France),
(*****) EXALEAD, 10 place de la Madeleine, 75008 Paris (France).

Mots clefs :
filtrage sémantique, indexation, recherche d'information
Keywords:
semantic screening, indexing, information retrieval
Palabras clave :
filtrado semántico, ajuste, busca d'información

Résumé

Actuellement développé et testé avec le soutien financier de l'ANR, FILTRAR-S est un outil d'analyse sémantique automatique de textes écrits qui combine les fonctions de filtrage, d'indexation et de fouille sur les documents indexés. Le filtrage est réalisé par l'extraction inductive de structures sémantiques associatives et conduit à l'indexation thématique du contenu des documents. La fouille du contenu des textes prend en compte une dimension factuelle par l'extraction non supervisée des relations entre entités nommées. L'interrogation en langage naturel est prise en charge par un système de question-réponse qui utilise les indexations thématiques et factuelles des textes produites par le système. Pour le filtrage, un objectif de modélisation du fonctionnement associatif de la mémoire sémantique nous a conduit à utiliser l'algorithme LDA du Topic Model. Les résultats d'une première expérimentation pour le traitement d'un corpus d'articles du journal « Le Monde » montre l'efficacité du système pour l'extraction des topics et l'indexation par topics des documents du corpus. Du point de vue de la fouille fondée sur la dimension factuelle des textes, une première expérimentation sur un corpus d'articles du « New York Times » donne également des résultats intéressants. Les développements en cours visent à mettre en œuvre des procédures de calcul de similarité sémantique entre le contenu d'un texte et celui de profils thématiques et factuels, ainsi qu'un fonctionnement interactif des modules, notamment en réponse à une question précise de l'utilisateur. Il faut enfin souligner que si FILTRAR-S est d'abord développé à des fins de sécurité et de protection du citoyen, les fonctionnalités dont il se dote pour la recherche d'information et pour la veille technologique intéressent des domaines divers.

1 Problématique

Dans cet article, nous présentons FILTRAR-S, un outil de traitement de l'information non structurée visant à apporter un soutien à une partie du processus de veille tel qu'il a été notamment défini par sa normalisation AFNOR [26]. Plus spécifiquement, FILTRAR-S combine des fonctionnalités de filtrage d'information et de fouille de textes s'appliquant à des documents issus de sources très diverses tels que ceux collectés à partir du Web. La spécificité de FILTRAR-S peut à cet égard se définir par la conjonction de quatre caractéristiques :

- il met en œuvre un filtrage s'appuyant sur un modèle de mémoire sémantique ;
- il associe filtrage et fouille de textes en combinant l'extraction de structures thématiques et la sélection d'information ;
- il offre des fonctionnalités de fouille de textes reposant sur des processus avancés de recherche d'information ;
- il associe étroitement les dimensions thématiques et factuelles dans ses fonctionnalités de fouille de textes.

Les deux premières caractéristiques sont à associer étroitement : le modèle proposé permet de définir chaque profil par la conjonction de thèmes induits à partir de corpus représentatifs des thématiques visées. Chaque profil est ainsi caractérisé par un ensemble de représentations signifiantes fondées sur des associations lexicales. Ces représentations sont intéressantes non seulement pour le filtrage proprement dit, mais également pour indexer les documents filtrés dans la perspective d'une fouille de nature thématique. Dans FILTRAR-S, une telle fouille peut en effet s'appuyer sur des requêtes adressées à un moteur de recherche « traditionnel » indexant les documents filtrés, mais peut aussi prendre en compte les thèmes associés à ces documents lors du filtrage.

FILTRAR-S offre en outre la possibilité – c'est le quatrième point ci-dessus – d'associer une telle fouille thématique à une fouille de nature plus factuelle. Celle-ci peut prendre deux formes complémentaires. Lorsque les informations recherchées sont très précises, cette fouille s'appuie sur un système de question-réponse, qui représente le pendant pour la dimension factuelle du moteur de recherche « traditionnel » pour la dimension thématique. Lorsque l'objectif de la fouille est plutôt d'avoir un panorama des relations entretenues par différentes entités, elle prend la forme d'un module d'extraction d'information non supervisée réalisant l'extraction de ces relations et leur regroupement suivant des critères de proximité sémantique et thématique.

Dans la suite de cet article, nous commencerons par donner une vue d'ensemble de FILTRAR-S et de ses fonctionnalités, puis nous décrirons les derniers développements réalisés concernant plus particulièrement deux de ses fonctionnalités les plus spécifiques : le filtrage et la fouille de textes centrée sur les relations entre entités.

2 Vue d'ensemble de FILTRAR-S

Les principes énoncés ci-dessus se concrétisent sous la forme de l'architecture illustrée par la figure 1. Son point d'entrée est constitué par les documents à l'issue de leur collecte et du traitement visant à extraire leur contenu textuel significatif. Dans l'outil FILTRAR-S, ces deux fonctionnalités sont assumées par un composant déjà existant développé par Exalead. Les documents ainsi produits font ensuite l'objet d'un traitement linguistique dont les résultats sont mis à disposition à la fois du module de filtrage et du module d'extraction d'information non supervisée. Ce traitement consiste à normaliser les mots par le biais de leur lemmatisation, à distinguer les mots grammaticaux des mots pleins et enfin, à identifier les entités nommées ayant un intérêt dans le domaine concerné. Les documents ainsi analysés passent par le module de filtrage qui ne retient que ceux en accord thématique avec les profils actifs. Dans le même temps, les documents filtrés sont indexés à la fois par les profils ayant conduit à leur sélection mais également par les thèmes associés à ces profils. Le module de filtrage sera plus spécifiquement abordé dans la section 3.

Outre cette indexation thématique, les documents filtrés sont indexés par les entités nommées qu'ils contiennent et les relations entre ces entités. Pour ce faire, un module d'extraction d'information non supervisée assure l'extraction de ces relations en s'appuyant sur les couples d'entités nommées, identifiées par le traitement linguistique initial, présents dans une même phrase. Dans une perspective de synthèse d'information pour la fouille de textes, ce module identifie également les

regroupements de relations équivalentes ainsi que les regroupements de relations se raccrochant à un même thème. Toutes ces informations d'indexation sont associées aux documents sous la forme d'annotations et conservées avec ceux-ci dans une base de textes servant de source aux processus de fouille textuelle.

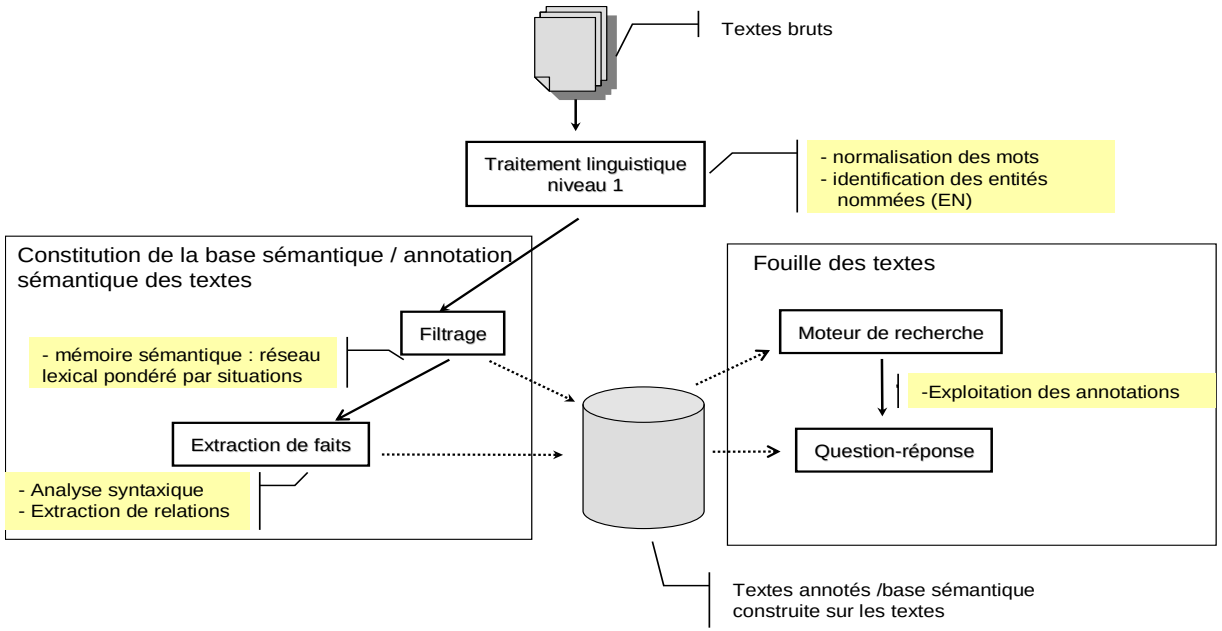

Figure 1 : Architecture de FILTRAR-S

Plus précisément, l'accès au contenu de cette base par ces processus de fouille s'effectue par le biais d'un moteur de recherche. Ce moteur est utilisé directement pour la fouille de nature thématique. Dans le cas de la fouille de nature factuelle, il ne constitue en revanche que le point d'accès au contenu de la base textuelle. Pour le système de question-réponse, il fournit classiquement les documents constituant le point de départ de la recherche des réponses. Néanmoins, dans FILTRAR-S, cette recherche peut aussi s'appuyer sur les annotations associées aux documents relatives aux entités nommées et aux relations entre entités en appariant les résultats de l'analyse des questions aux relations identifiées dans les documents et aux regroupements de relations équivalentes. Dans le cas de la recherche « orientée relation », c'est-à-dire de la recherche des relations unissant des entités données, l'exploitation de la base textuelle se focalise plus exclusivement sur les annotations produites par le module d'extraction d'information non supervisée, à la manière de Marchisio et al. [27], afin non seulement de fournir aux veilleurs les relations cherchées mais également de les présenter sous forme de regroupements suffisamment synthétiques et significatifs pour en faciliter l'appréhension.

La description de cette architecture montre l'importance, sur le plan de son implémentation, de disposer d'un formalisme commun de représentation des documents et de leurs annotations. Le choix s'est porté sur la plate-forme WebLab, qui présente l'avantage d'avoir été utilisée avec succès dans le cadre de plusieurs projets et qui se conforme aux standards du Web Sémantique. Les modules de traitement y apparaissent en effet sous la forme de services Web et le format de représentation des documents et de leurs annotations se fonde sur RDF.

3 Le module filtrage de FILTRAR-S

3.1 Les principes de l'approche

L'objectif du module est de réaliser un filtrage sémantique des informations contenues dans les textes. Il s'agit donc d'opérer une analyse de la signification des textes et non pas d'utiliser une correspondance terme à terme avec une série de mots clés.
Pour cela, notre approche consiste à simuler le fonctionnement associatif de la mémoire sémantique humaine. Les sciences cognitives ont largement montré que l'interprétation d'un texte fait appel à nos connaissances des situations et des événements qui se déroulent habituellement dans le monde. D'autre part, des études expérimentales ont montré que le traitement sémantique d'un mot n'active pas seulement la signification de ce mot dans la mémoire des individus. Il y active aussi la signification des mots auxquels le mot traité est associé par le biais de connaissances sur le monde [2, 3, 4, 6]. Des travaux complémentaires ont montré que ces processus associatifs interviennent pendant la compréhension d'un texte. La convergence de ces associations vers un même thème permet la récupération des connaissances du monde correspondantes avec pour conséquence la sélection des significations contextuellement appropriées [24, 28] et la production d'inférences [1, 29].
Mais le problème qui a longtemps empêché toute modélisation opérationnelle de ces mécanismes est la nécessité d'introduire dans le système les grandes quantités de connaissances du monde qui sous-tendent les associations sémantiques entre mots et qui sont récupérées en retour, via le traitement des mots associés. Les choses ont changé avec l'arrivée des modèles statistiques exploitant la fréquence de cooccurrence des mots dans de larges corpus de textes. Leur force est d'exploiter la grande quantité de données réelles que constituent les comportements verbaux inscrits dans les textes et d'extraire ainsi des connaissances sous la forme de structures lexicales associatives. Certains de ces modèles ont été utilisés comme modèles du fonctionnement associatif de la mémoire sémantique et c'est sur l'un de ces modèles que nous nous appuyons pour la conception du module de filtrage.

3.2 Modèles statistiques permettant l'extraction de structures sémantiques associatives

À la suite du modèle vectoriel de Salton [7], une série de modèles ont exploité le comptage des cooccurrences des mots dans les corpus de textes pour permettre la récupération automatique de textes. Ceux qui intéressent le système FILTRAR-S sont ceux qui cherchent également à construire des structures sémantiques et modéliser le fonctionnement de la mémoire sémantique humaine [5]. En effet, simuler ce fonctionnement cognitif humain entraîne un parallélisme entre les inférences associatives humaines et les réponses du système, ce qui tend à résoudre les problèmes de pertinence et d'adaptation à l'utilisateur.
Le plus connu des modèles qui visent la simulation du fonctionnement cognitif est LSA ou Latent Semantic Analysis [8]. Ce modèle utilise une technique mathématique qui réduit le nombre de dimensions d'une matrice Mot X Document, ce qui a pour effet de sélectionner les cooccurrences les plus significatives du corpus et d'extraire ainsi des structures sémantiques latentes qui contraignent les associations entre les mots représentés. Ce modèle donne des résultats qui ont une certaine validité psychologique, mais le principal problème est que les dimensions de l'espace ne sont pas directement interprétables. Il est donc difficile de sélectionner des sous-espaces vectoriels dans lesquels les nombreux mots polysémiques de tout corpus prennent un sens plutôt qu'un autre.
D'autres modèles ont fait l'économie de cette phase de réduction du nombre de dimension de la matrice, qui n'est en fait qu'une heuristique sans légitimité psychologique. Dans le modèle HAL, Burgess et Lund [9] comptabilisent les cooccurrences locales entre les mots d'un texte au moyen d'une « fenêtre glissante » de 10 mots. Ils construisent par ce moyen une matrice mot X mot de très grande taille (70 000 X 70 000) et un espace vectoriel à 140 000 dimensions (car ils font aussi intervenir l'ordre dans lequel co-occurrent les mots). Comptabiliser ainsi les cooccurrences dans une fenêtre de taille limitée respecte le fait que les capacités de la mémoire de travail humaine limitent le nombre de mots dont les représentations lexicales (signifiant et signifiés) peuvent être maintenues simultanément actives. Cependant, à l'inverse de LSA, le modèle accorde une grande importance aux cooccurrences locales et il exige de grandes capacités de mémoire pour gérer le grand nombre de dimensions de l'espace.

Une autre approche est celle du random indexing que Jones & Mewhort [10] utilisent dans le cadre de leur modèle Beagle. Trois vecteurs sont utilisés pour représenter les mots dans un espace vectoriel dont le nombre de dimension est fixé à priori. Un premier vecteur est censé coder la « forme perceptive » du mot et en l'état actuel du modèle, ses valeurs sont fixées aléatoirement. Par ailleurs, les valeurs de ce premier vecteur sont fixées de façon définitive. Un second vecteur est un vecteur « contexte » qui est recalculé à chaque occurrence du mot en faisant la somme des vecteurs perceptifs de chacun des mots qui co-occurrent avec ce mot. Sont comptabilisés comme co-occurrents tous les mots qui apparaissent entre deux signes de ponctuation successifs. Le troisième vecteur est un vecteur position qui comptabilise la distance moyenne avec laquelle chaque mot entre en cooccurrence avec ses voisins. Séduisant par son apprentissage contextuel cumulatif, ce modèle donne pour l'instant des performances équivalentes à celles du modèle LSA.

Le modèle dont s'inspire finalement le module de filtrage de FILTRAR-S est le Topic Model de Griffith, Steyvers et Tenenbaum [11]. Son principal avantage est d'extraire d'un corpus des topics, c'est à dire des groupes de mots associés qui relèvent d'un même contexte sémantique et contribuent à l'évocation d'un même thème ou une même situation. En outre, cette extraction de topics s'accompagne d'une indexation pondérée des textes du corpus à chacun de ses topics, ce qui constitue une première étape du filtrage.

Le Topic Model appartient à un courant de recherche qui applique les statistiques bayésiennes à la découverte de la sémantique des textes et qui a été initié par Hofmann [23]. Ce modèle génératif représente la signification des textes comme une série de topics sélectionnés en fonction de leur distribution probable en mémoire sémantique, et représente la forme de surface du texte comme une suite de mots sélectionnés en fonction de leur distribution de probabilité au sein des topics. Appliqué à un corpus de texte, le but du modèle est donc de déterminer la meilleure distribution de topics et de mots dans les topics, celle qui est la plus probable étant donné la fréquence de cooccurrence des mots constatés dans le corpus. Le modèle utilise l'algorithme LDA (Latent Dirichlet Allocation) de nature bayésienne non paramétrique [12]. LDA considère la distribution des mots dans un topic donné comme un échantillon d'une distribution de Dirichlet qui est elle-même la « conjugate prior » de la distribution multinomiale représentant la distribution des mots dans un topic. De la même façon, LDA considère la distribution des topics dans un document donné comme une autre distribution de Dirichlet, « conjugate prior » de la distribution multinomiale représentant la distribution des topics dans un document.

À partir d'une distribution multinomiale pour la fonction « likelihood » et d'une distribution de Dirichlet pour la « conjugate prior », on obtient la distribution a posteriori des variables recherchées. Dans le cadre du "Topic Model", les variables observées sont les occurrences des mots dans les documents et les variables recherchées sont les affectations de chacun des mots à un topic. L'échantillonnage de Gibbs, un algorithme de type MCMC (Markov chain Monte Carlo), permet à partir des mots observés de découvrir progressivement les topics apparaissant dans les documents. Les chercheurs ont montré que la distribution des mots dans les topics était une chaîne de Markov ergodique. Après quelques milliers d'itérations, la distribution des mots parmi les topics atteint un état stationnaire.

Le principal avantage du Topic Model est que les topics extraits sont des structures sémantiques directement interprétables. De ce fait, il est aisé de sélectionner une série de topics pour filtrer les documents d'un corpus qui traitent de ces topics et de ceux-là seulement. Des procédures de sélection de topics spécialisés extraits par LDA et de calcul de similarité sur ces sous-espaces ou bases sémantiques pourront alors être évaluées dans le cadre du système FILTRAR-S. En effet, une fois extraits à partir d'un corpus, procédure qui est gourmande en temps de calcul lorsque le corpus est de grande taille, les topics peuvent être assimilés aux dimensions d'un espace vectoriel. Cette possibilité a d'ailleurs été signalée par les auteurs du modèle [13], mais à notre connaissance elle n'a pas encore été véritablement appliquée. Par calcul de similarité elle devrait permettre le filtrage et la récupération de textes qui n'étaient pas présents dans le corpus d'apprentissage, mais dont le contenu est similaire à des textes types. Une procédure interactive de sélection des topics par proposition de mots associés à l'utilisateur est également à l'étude afin de permettre l'enrichissement des requêtes et la sélection des textes pertinents dans un topic. Ces procédures complémentaires devraient permettre au module de filtrage du système FILTRAR-S de répondre aux besoins d'utilisateurs d'une part pour le filtrage sémantique des textes en fonction veille et d'autre part pour la recherche d'information. Pour l'heure, ces procédures complémentaires ne sont pas encore opérationnelles, mais nous avons déjà franchi une étape importante qui est la réalisation d'un système capable d'extraire des topics et d'indexer les textes du corpus analysés à chacun de ces topics. Les résultats d'une première mise en œuvre du système sont présentés dans le paragraphe qui suit.

3.3 Premiers résultats du module filtrage : extraction de topics et classification des documents par topics

Cette première mise en œuvre a consisté à analyser le corpus du journal « Le Monde » pour l'année 2006. L'objectif était d'identifier les différents topics développés dans un large corpus de textes et de classer ces textes en fonction de leurs topics dominants.
Le corpus est fourni sous forme d'un document XML par article. Après un pré-traitement des articles (suppression des mots vides, lemmatisation), nous avons passé les articles dans une mise en œuvre LDA avec échantillonnage de Gibbs, Gibbs++.

3.3.1 Paramétrage

Pour le lancement de l'outil GibbsLDA++, les paramètres suivants ont été utilisés pour cette première expérimentation :

- Alpha : 0,166667 et Beta : 0,1
- nombre de documents : 43 627
- nombre de mots après lemmatisation : 208 895
- nombre de topics recherchés : 300
- nombre de topics filtrés : 126
- nombre d'itérations : 3000
- distribution minimum des mots dans le topic : 0,001
- nombre de mots dans le topic : 20
- nombre de documents dans le topic : 50

3.3.2 Résultats

Pour chacun des 126 topics filtrés, on a obtenu une distribution de probabilité pour 20 mots associés et on a récupéré 50 textes du corpus dont le contenu se rapporte principalement au topic, ainsi que la distribution de probabilité d'appartenance de chacun de ces textes au topic. Un échantillon des résultats obtenus est présenté dans les tableaux 1 et 2. Il s'agit du topic n° 12 dont le thème ou contexte sémantique est « la violence, les jeunes, la police, les banlieues... », ainsi que les mots du topic l'expriment clairement et plus complètement (cf. tableau 1). En outre, les titres des 50 textes sélectionnés pour ce topic, avec leur probabilité d'appartenance au topic (cf. tableau 1) permettent de constater que la plupart d'entre eux sont explicitement pertinents par rapport au topic.

Tableau 1 : Les mots du topic n° 12 et leur probabilité d'appartenance au topic

police	0,0303	policier	0,0290
violence	0,0205	jeune	0,0199
quartier	0,0117	banlieue	0,0098
cité	0,0086	intérieur	0,0070
sécurité	0,0058	délinquance	0,0051
urbain	0,0049	incident	0,0049
bande	0,0044	émeute	0,0043
voiture	0,0042	ordre	0,0041
gendarme	0,0039	Seine-st-Denis	0,0038
nuit	0,0036	incendie	0,0034

Tableau 2 : Les titres des 50 documents du topic n° 12 et leur probabilité d'appartenance au topic

1 [0,4658] Des bouteilles de gaz avec des clous retrouvées à Aulnay-sous-Bois
2 [0,4090] Un premier rapport de l'IGPN écarte la responsabilité des policiers
3 [0,4042] Des bandes affrontent la police dans le Val-d'Oise
4 [0,3985] Nouveaux incidents entre jeunes et policiers dans les Yvelines
5 [0,3866] Violentes émeutes après la mort de deux adolescents dans une collision avec la police
6 [0,3840] La police trouve des armes à feu après un affrontement à Neuilly-Plaisance
7 [0,3651] Bagarres à répétition entre bandes rivales dans le quartier de la gare du Nord à Paris
8 [0,3639] A Paris, la vendetta entre GDN et Def Mafia déborde dans la rue
9 [0,3527] Les chiffres de la police surestiment le phénomène des bandes
10 [0,352684] Des émeutes urbaines à Cergy sont passées inaperçues
11 [0,3477] Violences urbaines à Saint-Dizier
12 [0,3338] Des gardiens de HLM manifestent leur ras-le-bol après des agressions
13 [0,3310] Les quartiers nord d'Aulnaysous surveillance par peur de nouveaux affrontements
14 [0,330 Emeutes : pourquoi 2007 diffère de 2005
15 [0,3290] Quatorze personnes en garde à vue après des bagarres entre bandes rivales à Paris
16 [0,3283] Provocations policières
17 [0,3197] Près de 730 voitures ont été brûlées au cours de la nuit suivant l'élection
18 [0,3195] Des bandes rivales et la police se sont affrontées à Cergy sans faire événement
19 [0,3167] Treize personnes mises en examen dans l'enquête sur les émeutes de Saint-Dizier
20 [0,31263] L'adolescent renversé par une voiture de police samedi à Marseille est mort
21 [0,3085] Un jeune de 15 ans interpellé dans l'affaire de la mort d'un policier
22 [0,3073] Les bandes sous la loupe des RG
23 [0,3069] Les violences antipolice, nouveau tabou
24 [0,3061] Aulnay-sous-Bois : incidents entre jeunes et policiers
25 [0,3024] Une vidéo montre l'arrestation violente de deux hommes près de Rouen
26 [0,2988] Guerre de tranchées dans le Val-d'Oise
27 [0,2974] Selon l'IGPN, la voiture aurait été dégradée après l'accident
28 [0,29573] Une vidéo témoigne de l'état de la voiture de police après l'accident
29 [0,2953] Policiers et groupes de jeunes se sont affrontés gare du Nord, à Paris
30 [0,2947] Les femmes sont victimes plus que les hommes de la violence
31 [0,287475] La tactique de la police face à des « groupes hostiles très mobiles »
32 [0,287155] Une lettre du préfet du Val-d'Oise
33 [0,286693] Retour sur une émeute
34 [0,278700] Un « effet de foule » évoqué dans la mort d'un policier
35 [0,273933] Des salariés menacés de licenciements ont été empêchés de se rendre à un meeting
36 [0,265088] Dans l'Essonne, la tension ordinaire entre jeunes et policiers
37 [0,263468] Deux policiers condamnés pour des violences à Saint-Denis
38 [0,262160] Dans les cités d'Ile-de-France, des jeunes gens craignent la victoire du « candidat de la police »

39 [0,262037] Bagarre de la rue Lafayette : simple « rixe », selon les juges
40 [0,261582] Quand la police use de la force, c'est qu'elle est faible »
41 [0,258370] Un policier en service meurt écrasé par un manège à Paris
42 [0,257106] Pour commencer l'année sur le câble et le satellite
43 [0,253388] Marche silencieuse en hommage à Larami Samoura, l'un des deux adolescents morts à Villiers-le-Bel
44 [0,252210] Précaire retour au calme après les émeutes de Villiers-le-Bel
45 [0,251589] Il n'y a pas eu d'émeutes urbaines à Cergy
46 [0,249507] Nouvelle relaxe d'un jeune poursuivi pour l'incendie d'un bus à Grigny en 2006
47 [0,249308] A Toulouse, la police en mal de proximité et de résultats
48 [0,248831] Banlieues : scènes de guérilla urbaine à Villiers-le-Bel
49 [0,246683] Un inconnu nommé Jean Moulin
50 [0,245505] Une nuit « calme » à Villiers-le-Bel

On constate que pour le topic présenté dans le tableau 1, seuls les textes 30, 42 et 49 sont non pertinents par rapport au topic, ce qui équivaut à un taux d'erreurs de 6%. Par ailleurs, une analyse de l'ensemble des 126 topics extraits du corpus nous a permis de constater que 4 d'entre eux rassemblent des mots qui ne permettent pas de dégager un thème cohérent et interprétable. Quatre autres topics rassemblent des mots qui appartiennent visiblement à deux thèmes indépendants l'un de l'autre. Par exemple, le topic n° 14 contient les mots [tabac fumer cigarette fumeur descente ski alpin slalom autrichien skieur …].
On estime d'autre part que 38% des topics extraits expriment des thèmes généraux. Par exemple, le topic n° 108 contient les mots : [université école étudiant élève éducation enseignant enseignement scolaire professeur …]. À l'inverse, 13,5% des topics concernent un fait unique et précis qui a défrayé la chronique. Par exemple, le topic n° 29 concerne exclusivement l'affaire Maddie McCann et contient entre autres les mots [portugais McCann portugal Maddie Madeleine Kate Gerry disparition ...].
Ces variations estimées du degré de généralité des topics pour un même paramétrage des algorithmes révèlent l'influence de la composition du corpus traité et de la redondance du contenu des textes qu'il contient. Ces éléments doivent être pris en compte pour le paramétrage de l'algorithme en fonction des objectifs et devraient faire l'objet de tests pour optimiser le rendement de l'algorithme. Griffith et Steyvers [25] soulignent à ce propos qu'une augmentation de la valeur du paramètre Béta augmente la granularité des topics, tout comme le nombre de topics recherchés. Dans leur application qui concernait la classification par topic de 28 154 articles scientifiques à partir de leurs résumés, Griffith et Steyvers ont testé jusqu'à 1000 topics, mais estiment que le chiffre de 300 topics est optimal.

3.3.3 Conclusions de l'expérimentation

La conclusion de notre première expérimentation du module est que les algorithmes utilisés sont efficaces, mais que différents paramétrages doivent maintenant être testés afin que la granularité des topics extraits soit optimale par rapport aux objectifs de l'utilisateur. En outre, il nous faut maintenant mettre en œuvre les procédures complémentaires de calcul de similarité et d'expansion de requête. Ainsi le module de filtrage pourra être utilisé à la fois pour filtrer les documents extérieurs au corpus d'apprentissage dans le cadre d'une fonction veille, et pour enrichir de mots clés pertinents les requêtes d'utilisateurs dans le cadre d'une application de type moteur de recherche, pour enrichir des requêtes d'utilisateur.

4 Extraction d'information non supervisée

Dans cette section, nous aborderons la dimension factuelle de la composante « fouille de textes » de FILTRAR-S en rendant compte des travaux que nous avons réalisés dans le domaine de l'extraction d'information non supervisée. Nous commencerons par une présentation rapide de ce champ de recherche encore relativement neuf avant de rendre compte de nos premiers résultats le concernant.

4.1 Principes

L'extraction d'information, popularisée notamment par les conférences MUC (Message Understanding Conferences) [15], consiste classiquement à repérer dans les textes des événements d'un type prédéfini ainsi qu'un ensemble donné d'informations prenant généralement la forme d'entités nommées et venant s'insérer dans une description *a priori* de ce type d'événements appelée *template*. Pour un événement comme le rachat d'une société par une autre, l'extraction se focalisera ainsi sur l'identification de la société acheteuse, de la société achetée, du montant du rachat et de sa date. Cette approche peut être qualifiée globalement de dirigée par les buts ou de supervisée. Plus récemment, une approche inverse a fait son apparition, approche que nous qualifierons ici d'extraction d'information non supervisée ([16, 17, 18]). Cette approche prend comme point de départ des entités ou des types d'entités et se fixe comme objectif de mettre en évidence les relations intervenant entre ces entités puis de regrouper ces relations en fonction de leurs similarités sémantiques ou thématiques. Une telle approche s'incarne typiquement dans une problématique de veille telle que « suivre tous les événements faisant intervenir les sociétés X et Y ».
En complément des capacités de recherche thématique offertes par un moteur de recherche « classique » et de la possibilité de répondre à des questions factuelles, FILTRAR-S propose également des fonctionnalités d'extraction d'information non supervisée afin qu'un utilisateur puisse explorer les relations intervenant entre des entités ou des types d'entités et en avoir une vue structurée. Par ailleurs, les relations ainsi mises en évidence sont également indexées pour être exploitées afin de répondre à des questions.

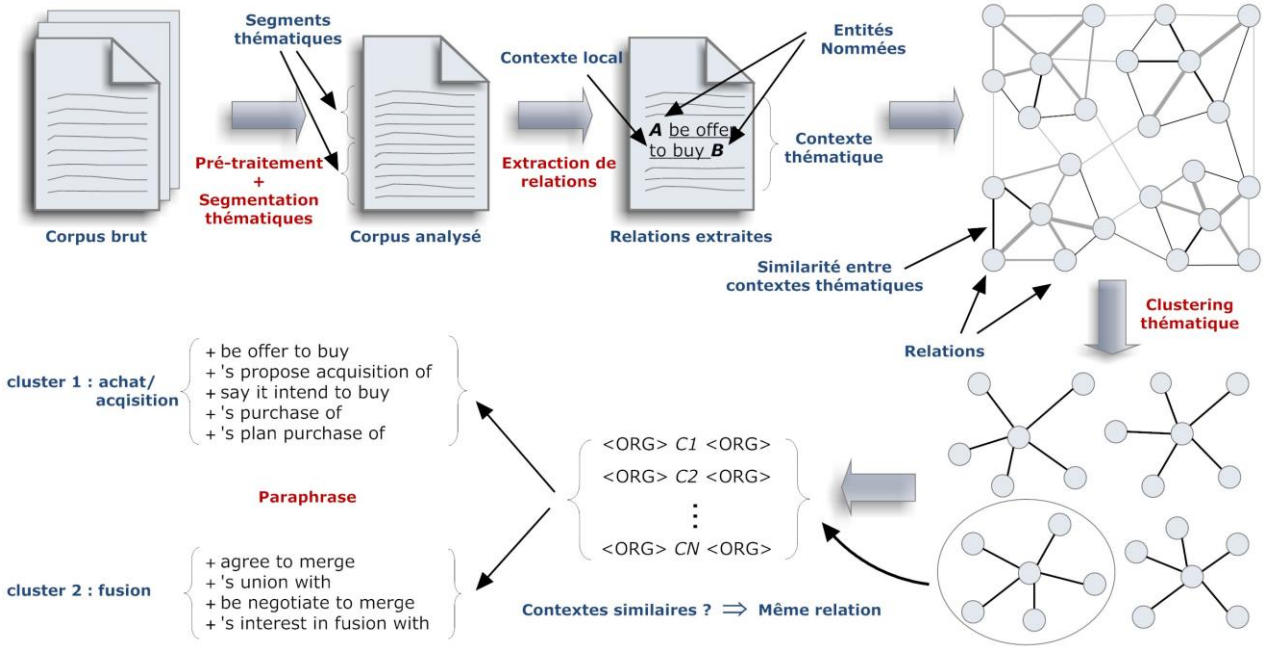

Figure 2 : Principes de l'expérimentation menée

Dans la perspective de mettre en œuvre ces nouvelles capacités de recherche, nous avons réalisé une première expérimentation (cf. figure 2) couvrant à la fois l'extraction de relations et leur structuration sur un plan sémantique et thématique.

4.2 Extraction de relations entre entités

Dans le contexte de ce travail, une relation, ou plus précisément une instance de relation, se définit comme un couple d'entités nommées extraites d'une phrase. Outre la donnée de ce couple d'entités, une telle instance de relation est définie par deux éléments :

- une caractérisation linguistique de la relation sous-jacente au couple d'entités considéré. En l'occurrence, cette caractérisation distingue trois composantes : le segment de phrase entre les deux entités (*in*), le segment de phrase précédant la première entité (*bf*) et celui suivant la seconde (*af*). Cette définition linguistique de la relation est le support du regroupement des instances de relations sémantiquement similaires ;
- une caractérisation thématique, appelée contexte thématique (*ct*), formée des mots du segment de texte dont l'instance de relation considérée a été extraite. Ces mots sont pondérés suivant la méthode *tf.idf*, adoptée classiquement en recherche d'information pour rendre compte de l'importance des termes par rapport à un corpus. Cette caractérisation est le point de départ du regroupement thématique des instances de relation.

La figure 3 illustre cette représentation des relations en donnant sa forme XML pour la relation *Adrienne Redd — Cabrini College* extraite de la phrase : *« I am agonizing over this election, » said Adrienne Redd, 43, who teaches at Cabrini College and has two children*. Pour des raisons d'espace, seuls quelques mots de la caractérisation thématique de la relation sont donnés. Chaque mot apparaissant dans cette représentation est caractérisé par sa forme fléchie (*inf*), son lemme (*lem*), son stemme (*stem*) et sa position (*pos*).

La mise en œuvre de l'extraction de telles instances de relations s'appuie sur trois traitements appliqués aux textes :

- la reconnaissance des entités nommées. Dans le cas présent, nous nous sommes focalisés sur les entités nommées générales distinguées dans le cadre de MUC : personnes, organisations, lieux, dates, unités temporelles, pourcentages et unités monétaires ;
- la lemmatisation. Celle-ci conduit à normaliser les mots des textes et passe par la détermination de leur catégorie grammaticale. Elle permet ainsi de distinguer les mots grammaticaux des mots porteurs de contenu informationnel, dits mots pleins. Cette lemmatisation est en particulier appliquée à la caractérisation linguistique des relations ;
- la segmentation thématique. Cette segmentation découpe les textes en segments thématiquement homogènes. Chaque instance de relation se voit ainsi associer les mots pleins normalisés du segment thématique dans lequel elle se trouve.

Dans le cadre de l'expérimentation dont les résultats sont rapportés à la section 4.4, nous avons utilisé les outils applicables à l'anglais d'OpenNLP [19] pour la lemmatisation et la reconnaissance des entités nommées. Concernant la segmentation thématique, nous nous sommes appuyés sur l'outil LCseg [20].

Figure 3 : Exemple de représentation d'une relation entre une personne et une organisation

4.3 Structuration des relations extraites

Ainsi que l'illustre la figure 2, les instances de relations extraites des textes sont structurées par le biais d'une double étape de regroupement. Le premier regroupement se focalise sur les contextes thématiques des relations. Son objectif est de rassembler les instances de relations intervenant dans un même cadre thématique, une même situation ou relatives à un même événement. Il est mis en œuvre en appliquant l'algorithme *Markov Clustering* [21]. Cet algorithme, de type *Random Walk*, simule plusieurs écoulements aléatoires dans un graphe, renforce l'écoulement là où il est fort et l'affaiblit là où il est faible. Dans le cas présent, son

point de départ est le graphe de similarité des contextes thématiques, obtenu en calculant la mesure *cosinus* deux à deux entre les représentations vectorielles des contextes thématiques des instances de relations. Lorsque le nombre de contextes thématiques est important, il est possible de faire appel à des algorithmes permettant d'obtenir les contextes les plus proches sans avoir à calculer toutes les similarités. Pour notre part, nous nous appuyons sur l'algorithme *All Pairs Similarity Search* proposé dans [22]. Cet algorithme permet d'obtenir efficacement tous les couples d'objets considérés, ici les contextes thématiques des relations, dont la similarité est supérieure ou égale à un seuil fixé *a priori* selon la mesure *cosinus*.

Le second regroupement s'effectue quant à lui au sein de chacun des groupes thématiques de relations. Le même algorithme, *Markov Clustering*, est alors appliqué aux caractérisations linguistiques des instances de relations regroupées afin d'identifier les instances de relations sémantiquement similaires, c'est-à-dire pouvant être considérées comme des paraphrases. Dans ce cas, deux mesures de similarité ont été testées : la mesure *cosinus*, permettant une comparaison de type « sac de mots », et la distance d'édition, tenant compte de l'ordre des éléments linguistiques présents dans les caractérisations de relations. Celles-ci étant constituées de trois parties, différentes combinaisons de ces parties ont en outre été testées.

4.4 Résultats et évaluation

L'approche proposée a été expérimentée sur une sous-partie du corpus AQUAINT-2 constituée de 18 mois du journal *New York Times*. Le tableau 3 donne un aperçu de la volumétrie des relations extraites en se restreignant aux entités de type personne, lieu et organisation. Le tableau fait apparaître plus précisément le nombre brut de relations extraites et le nombre de ces relations après dédoublonnage. Certaines relations sont en effet identiques au niveau de leur caractérisation linguistique du fait de la répétition de certaines phrases dans le corpus. Ces phrases répétées se retrouvent lorsque plusieurs articles relatifs à un même sujet sont très proches ou dans des rubriques très formatées. Le dédoublonnage est réalisé selon la procédure suivante :

- transformation des relations en une représentation de type « sac de mots » en se limitant aux éléments de leur caractérisation linguistique. Les entités nommées de la relation sont représentées de façon spécifique en concaténant leurs constituants, leur type et leur position dans la relation ;
- sélection des couples de relations dont la similarité des représentations selon la mesure *cosinus* est égale à 1,0 en s'appuyant sur l'algorithme *All Pairs Similarity Search* ;
- utilisation de l'algorithme *Markov Clustering* pour regrouper les relations similaires ;
- choix d'une relation au sein de chaque regroupement formé pour représenter toutes les relations du regroupement.

Tableau 3 : Volumétrie des relations extraites

Relation	Nombre de relations extraites	Nombre de relations après dédoublonnage
PERSON – PERSON	175 802	160 562
PERSON – ORGANIZATION	126 281	108 157
PERSON -LOCATION	152 514	134 768
ORGANIZATION – ORGANIZATION	77 025	66 902
ORGANIZATION – LOCATION	71 858	64 432
ORGANIZATION – PERSON	73 895	65 603

Une première expérimentation de regroupement thématique de relations à large échelle a été réalisée pour les 160 562 relations intervenant entre des entités de type personne. En l'absence de référence, il est difficile de caractériser le résultat d'un tel clustering en termes de pertinence mais la distribution de la taille des regroupements formés donnée par le tableau 4 permet de voir que si l'on observe le schéma assez classique d'un large ensemble de petits regroupements accompagné d'un nombre limité de gros regroupements, la distribution entre ces deux extrêmes est néanmoins peuplée de façon assez équilibrée. On peut donc faire l'hypothèse qu'une partie significative des regroupements formés correspondent à des thématiques à la fois ni trop générales, ni trop spécifiques. Cette hypothèse reste néanmoins à confirmer au travers d'un examen plus approfondi des regroupements à un niveau global.

Tableau 4 : Distribution des regroupements thématiques de relations entre des entités de type personne

Taille des regroupements	1	2 -3	4 -10	11 - 100	> 100
Nombre de regroupements	33	25 208	9 321	871	8

Une évaluation plus précise des étapes de regroupement a cependant été menée et s'est faite sur un sous-ensemble très restreint de ces relations du fait de la nécessité d'un jugement humain *a posteriori* de la validité des regroupements effectués. Plus précisément, cette évaluation a été réalisée sur le sous-ensemble des 718 instances de relations entre personnes dont l'une des entités était « Dick Cheney ». Ces instances ont été réparties en deux grands clusters thématiques par la première étape de regroupement et au sein de chacun d'entre eux, plusieurs dizaines de clusters sémantiques ont été formés par la seconde étape de regroupement. L'évaluation de la pertinence de ces regroupements finaux a été réalisée en sélectionnant aléatoirement 50 instances de relations et en jugeant pour chaque couple de relations la pertinence du regroupement ou du non-regroupement de ces instances. À l'instar de [18], nous avons utilisé la racine carrée du coefficient de Jaccard comme mesure d'évaluation globale de la pertinence d'un regroupement *R* par rapport à un regroupement de référence *M* :

où *SS* est le nombre d'instances classées ensemble dans *R* et *M*, *SD*, le nombre d'instances classées ensemble dans *R* mais pas dans *M* et *DS*, le nombre d'instances classées ensemble dans *M* mais pas dans *R*. Nous avons aussi utilisé l'*Accuracy*, définie par :

avec *DD*, le nombre d'instances classées séparément dans *M* et *R*.

Tableau 4 : Évaluation des regroupements

mesure – taille caractérisation	JC	Accuracy	# clusters
cos-mid	0,24	0,89	105
cos-sent	0,30	0,78	41
ed-mid	0,28	0,48	32
ed-sent	0,31	0,36	2

Le tableau 4 montre les valeurs des mesures ci-dessus ainsi que le nombre de regroupements formés pour différentes combinaisons de mesures de similarité (cos : cosinus, ed : distance d'édition) et de taille de la caractérisation linguistique de la relation (mid : partie de phrase entre les deux entités ; sent : toute la phrase). Visiblement, la combinaison cosinus-phrase donne le meilleur résultat. Avoir une caractérisation linguistique plus courte conduit à un plus grand nombre de clusters tandis que l'approche « sac de mots » de la mesure cosinus semble plus effective que la prise en compte de la linéarité des énoncés faite par la distance d'édition.

5 Conclusion et perspectives

Dans cet article, nous avons présenté FILTRAR-S, un système associant filtrage d'information et fouille de textes. Du point de vue du filtrage, FILTRAR-S présente la particularité de s'appuyer sur un modèle de mémoire sémantique lui conférant une plus grande intelligibilité des résultats par le biais de l'utilisation de représentations lexicales de thèmes induits à partir de corpus. Du point de vue de la fouille des textes, FILTRAR-S offre des outils prenant en compte à la fois les dimensions thématique et factuelle des textes en s'appuyant dans ce dernier cas sur des outils de recherche d'information avancée comme les systèmes de question-réponse ou d'extraction d'information non supervisée. Nous avons présenté ici les premiers résultats de FILTRAR-S concernant le filtrage et la fouille des textes orientée par les relations entre entités. Dans le cas du filtrage, outre l'extraction d'un corpus initial des structures lexicales associatives que constituent les topics et l'indexation des textes du corpus en fonction de ces topics, une validation de l'utilisation des topics pour le filtrage de nouveaux documents dans le cadre d'une fonction veille est en cours. Du côté de la fouille de textes, outre l'extension des résultats déjà obtenus, un travail d'intégration plus étroite est encore à mener, en particulier afin que le module de question-réponse soit à même d'exploiter les annotations produites par le module d'extraction d'information non supervisée.
Plus généralement, bien que le système FILTRAR-S soit développé dans le cadre d'un projet du programme « Concepts, Systèmes et Outils pour la Sécurité Globale » (CSOSG) de l'ANR, nous considérons que son champ d'action s'étend bien au-delà de la sphère de la sécurité et qu'il peut tout à fait être utilisé comme outil de support d'un processus de veille dans des domaines tels que le domaine biomédical ou l'e-commerce par exemple.

6 Bibliographie

[1] CAMPION N., MARTINS D. and WHILHEM A., *Contradictions and Predictions: Two Sources of Uncertainty that Raise the Cognitive Interest of Readers*, Discourse Processes, 46, p 1–28, 2009

[2] HARE M., JONES M., THOMSON C., KELLY S. and MCRAE, K., *Activating event knowledge*, Cognition, 111 (2), p 151–167, 2009

[3] CAMPION N., ROSSI J.-P., LE NY J.-F. and DECLERCQ C., *Action schema, a basic knowledge structure accessed to provide meaning relations to words*, Sixteenth Annual Meeting of the Society for Text and Discourse, 13-15 juillet, Minneapolis, USA, 2006

[4] CAMPION N. et ROSSI J.-P., *Les schémas d'actions : structure et fonction d'un composant de la mémoire sémantique*, Communication orale au colloque de L'Arco, Lyon, 2008

[5] CAMPION N., CLOSSON J., CARCENAC J., FERRET O., GRAU B. et SHIN, J., *Modélisation de la mémoire sémantique et compréhension des messages par Filtrar-S : Un cyber-outil pour la sécurité globale*, Actes du troisième Workshop Interdisciplinaire sur la Sécurité Globale (WISG 2009), Troyes, France, 2009

[6] CAMPION N. et RIGALLEAU, F., *Les atteintes fonctionnelles précoces de la mémoire sémantique dans la maladie d'Alzheimer : Diagnostic et remédiation cognitive*, Colloque ARCo'08 - Connaissances : Genèse, Nature et Fonction, 3-5 décembre, Lyon, 2008

[7] SALTON G. and MCGILL M., *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983

[8] LANDAUER T. K. and DUMAIS S. T., *A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge*, Psychological Review, 104(2), p 211–240, 1997

[9] BURGESS C. and LUND K., *The dynamics of meaning in memory*, In Dietrich, E., & Markman, A. B. (Eds.), Cognitive Dynamics: Conceptual Change in Humans and Machines, Lawrence Erlbaum Associates Publishers, p 117–156, 2000

[10] JONES M. N. and MEWHORT D. J. K., *Representing Word Meaning and Order Information in a Composite Holographic Lexicon*, Psychological Review, 114(1), p 1–37, 2007

[11] GRIFFITHS T. L., STEYVERS M. and TENENBAUM J. B., *Topics in semantic representation*, Psychological Review, 114, p 211–244, 2007

[12] BLEI D.M., NG A. Y. and JORDAN M. I., *Latent Dirichlet allocation*, Journal of Machine Learning Research, 3, p 993–1022, 2003

[13] STEYVERS M. and GRIFFITHS T., *Probabilistic topic models*, In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), Handbook of Latent Semantic Analysis, Hillsdale, NJ: Erlbaum. p 427–448, 2007

[14] DUMAIS S. T., *LSA and information retrieval: Getting back to the basics*. In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), Handbook of Latent Semantic Analysis, Hillsdale, NJ: Erlbaum. p 293–321, 2007

[15] GRISHAM R. and SUNDHEIM B., *Design of the MUC-6 evaluation*. 6th conference on Message understanding, 1995.

[16] HASEGAWA T., SEKINE S. and GRISHMAN R., *Discovering Relations among Named Entities from Large Corpora*. 42nd Meeting of the Association for Computational Linguistics (ACL'04), p 415–422, 2004

[17] SHINYAMA Y. and SEKINE S., *Preemptive Information Extraction using Unrestricted Relation Discovery*, HLT-NAACL, p 304–311, 2006

[18] ROSENFELD B. and FELDMAN R., *Clustering for unsupervised relation identification*. Sixteenth ACM conference on Conference on information and knowledge management (CIKM'07), p 411 – 418, 2007

[19] OpenNLP : http://opennlp.sourceforge.net/

[20] GALLEY M., MCKEOWN K., FOSLER-LUSSIER E. and JING H., *Discourse Segmentation of Multi-Party Conversation*. 41st Annual Meeting on Association for Computational Linguistics, p 562–569, 2003

[21] VAN DONGEN S., *Graph clustering by flow simulation*. PhD thesis, University of Utrecht, 2000.

[22] BAYARDO R. J., MA Y. and SRIKANT R., *Scaling Up All-Pairs Similarity Search*. 16th International Conference on World Wide Web (WWW 2007), 2007.

[23] HOFMANN T., *Probabilistic latent semantic indexing*, 22nd Annual International SIGIR Conference on Research and Development in Information Retrieval, p 50–57, 1999

[24] TILL R., MROSS E. and KINTSCH W., *Time course of priming for associate and inference words in a discourse context*, Memory and Cognition, 16, p 283–298, 1988

[25] GRIFFITHS T. L. and STEYVERS M., *Finding scientific topics, National Academy of Sciences, 101*, p 5228–5235, 2004

[26] HERMEL L., *Maîtriser et pratiquer la veille stratégique*, AFNOR, 2001

[27] MARCHISIO G, DHILLON D, LIANG J, TUSK C, KOPERSKI K, NGUYEN T, WHITE D and POCHMAN L. *A Case Study in Natural Language Based Web Search*, In A. Kao and S.R. Poteet (Eds.), Text Mining and Natural Language Processing, Springer, 2006.

[28] KINTSCH W., *On the notions of theme and topic in psychological process models of text comprehension*, In W. van Peer and M.M. Louwerse (Eds.), Thematics in psychology and literary studies, p 157–170), Amsterdam: Benjamins, 2002

[29] CAMPION N. and ROSSI J.-P., *Associative and causal constraints in the process of generating predictive inferences*, Discourse Processes, 31(3), p 263–291, 2001