Extraction automatique d'information inattendue à partir de textes.

François Jacquenet, Christine Largeron

Université Jean Monnet de Saint-Etienne
EURISE
23 rue du docteur Paul Michelon
42023 Saint-Etienne Cedex 2
Francois.Jacquenet@univ-st-etienne.fr
Christine.Largeron@univ-st-etienne.fr

Résumé. Dans cet article, nous proposons d'utiliser des techniques de fouille de textes pour extraire des informations, automatiquement et à des fins stratégiques, à partir de bases de données scientifiques et techniques. Ce contexte de veille technologique introduit une difficulté inhabituelle par rapport aux domaines d'application classiques de la fouille de textes, puisqu'au lieu de rechercher de la connaissance fréquente cachée dans les données, il faut rechercher de la connaissance inattendue, qualifiée par les veilleurs de signal faible. Les mesures usuelles d'extraction de la connaissance à partir de textes doivent, de ce fait, être revues. Pour ce faire, nous avons développé le système *UnexpectedMiner* dans lequel de nouvelles mesures permettent d'estimer le caractère inattendu d'un document. Notre système est évalué sur une base de résumés d'articles scientifiques.

1 Introduction

Du fait de l'augmentation croissante des capacités de production et de stockage des données, des travaux ont porté tout d'abord sur le développement de méthodes et d'algorithmes permettant de les analyser et d'en extraire automatiquement des connaissances utiles. Mais rapidement, en plus du problème du volume des données c'est celui de leur diversité et de leur hétérogénéité qui a suscité l'intérêt des chercheurs. L'enjeu est en effet, de sortir du carcan tabulaire à n lignes (les individus) et p colonnes (les attributs associés à des types standards de données : booléen, nominal ou ordinal, réel) pour répondre aux besoins de nouvelles applications. Parmi les formes de données considérées on peut citer les données multi-relationnelles, issues de plusieurs tables d'une base de données relationnelles, qui ont conduit à la conception de Data Warehouses et de systèmes OLAP (*On-Line Analytical Procesing*). On peut aussi évoquer les données transactionnelles, où chaque enregistrement d'un fichier est formé d'un identifiant et d'un ensemble d'items (dont l'exemple type est le panier d'une ménagère) et, qui ont pu être traitées efficacement à l'aide d'algorithmes d'extraction de règles d'association. On peut mentionner également les données séquentielles décrivant l'évolution dans le temps d'un phénomène, et qui ont été étudiées par différentes communautés, par exemple en apprentissage automatique à l'aide d'automates ou en statistiques par des modèles stochastiques et régressifs. Sans parler de données encore plus complexes

que ce soit en raison de leur hétérogénéité lorsqu'elles combinent plusieurs types de données comme par exemple les données multi-media, ou en raison de leurs origines différentes comme les bases de données réparties.

Dans cet article, nous nous intéressons à des données textuelles et, plus particulièrement, à la recherche d'informations inattendues dans des textes scientifiques pour la veille technologique [Desvals et Dou, 1992, Jakobiak, 1990, Jakobiak, 1994]. La veille technologique consiste à rechercher et à exploiter à des fins décisionnelles des informations de nature scientifique et technique telles que des brevets, des articles scientifiques, ou encore des thèses. Le processus de veille peut être décomposé en quatre phases principales : l'audit des besoins, la collecte des données, le traitement de celles-ci et enfin, la synthèse et la diffusion des résultats. A chacune de ces étapes, le recours à des outils de traitements automatisés ou semi-automatisés paraît d'autant plus justifié que le volume de données disponibles est de plus en plus important et que celles-ci sont la plupart du temps accessibles sous forme numérique dans des banques de données ou par Internet. Il existe d'ailleurs déjà un certain nombre d'outils utilisables en intelligence économique. Une typologie en a été proposée récemment [François, 2003] en distinguant, en fonction de leur usage, les moteurs de recherche d'information et les agents intelligents, les logiciels de surveillance et de push, les systèmes de gestion de documents et de l'information et, les outils d'exploration du contenu des documents. Ces produits répondent cependant actuellement mal à l'attente des veilleurs, soit parce qu'il s'agit d'outils très spécialisés qui nécessitent un paramétrage assez lourd en fonction du sujet de veille, soit au contraire parce qu'il s'agit de logiciels généraux, comme les moteurs de recherche, qui supposent une connaissance a priori assez précise de ce que l'on recherche. Pour cette raison, les techniques de fouille de données paraissent un moyen adéquat pour automatiser la veille.

La fouille de données a en effet été définie comme "l'extraction non triviale, à partir de données, d'une information potentiellement utile, implicite et inconnue auparavant" [Fayyad *et al.*, 1996]. Elle a connu un fort développement depuis le milieu des années 90 du fait de la mise au point de nouveaux algorithmes performants permettant de traiter de gros volumes de données dans le domaine commercial. Lorsque les données considérées se présentent sous la forme de textes, qu'ils soient structurés ou non, on parle alors de fouille de textes (*text mining*). Par analogie avec la fouille de données, la fouille de textes [Kodratoff, 1999], introduite en 1995 par Feldman [Feldman et Dagan, 1995], est définie par Sebastiani [Sebastiani, 2002] comme l'ensemble des tâches qui, par analyse de grandes quantités de textes et la détection de modèles fréquents, essaie d'extraire de l'information probablement utile.
De fait, la fouille de textes constitue déjà un champ de recherche important qui a permis l'émergence de techniques utilisables dans le contexte de la veille scientifique et technique. Losiewicz et al. [Losiewicz *et al.*, 2000] par exemple, montrent comment les techniques de classification, de résumé et d'extraction de connaissances peuvent être utilisées à des fins décisionnelles. Zhu et Porter [Zhu *et al.*, 1999, Zhu et Porter, 2002] adoptent une approche bibliométrique pour détecter des opportunités technologiques à partir des informations concurrentielles relevées dans des documents électroniques.

Les travaux de Lent [Lent *et al.*, 1997] constitue un autre exemple d'application des techniques de fouille de textes en intelligence économique, dans lequel des algorithmes d'extraction de motifs séquentiels fréquents [Agrawal et Srikant, 1995] servent à déceler de nouvelles tendances dans une base de données de brevets chez IBM. Le principe consiste à observer des séquences de mots sur une période de temps de façon à repérer celles qui deviennent un motif fréquent alors qu'elles ne l'étaient pas initialement. De même, mais dans le cadre de la détection et du suivi de thèmes (*Topic Detection Tracking* [1]), Rajaraman *et al.* [Rajaraman et Tan, 2001] utilisent des réseaux de neurones pour découvrir des tendances à partir de bases de textes. Dans tous ces exemples, les techniques de fouille de textes sont principalement employées pour extraire, à partir de gros volumes de données, des informations utiles et fréquentes ou encore, pour identifier celles qui se rapportent à un sujet donné. Or, un des enjeux de la veille scientifique et technique, et plus généralement de l'intelligence économique, réside dans la détection d'informations nouvelles et inattendues. De telles informations, qualifiées d'ailleurs par les veilleurs de signaux faibles, n'apparaissent donc pas en général avec une fréquence élevée. De ce fait, les algorithmes d'extraction de motifs séquentiels fréquents, employés habituellement en fouille de données, semblent inappropriés dans le cadre de la veille puisque, comme leur nom l'indique, ils s'intéressent aux informations qui apparaissent fréquemment dans une base de données. C'est vraisemblablement une des raisons principales pour laquelle les logiciels commerciaux répondent mal actuellement à l'attente des veilleurs.

Partant de ce constat, un certain nombre de recherches se sont focalisées sur ce qui est appelé *évènement rare*, *information inattendue* ou encore *thème émergent*, selon les auteurs. Ainsi, Bun *et al.* [Bun et Ishizuka, 2001] ont proposé un système de détection de thèmes émergents. Ce système observe les changements qui interviennent sur un ensemble de sites Web et, il repère les mots apparaissant dans les pages modifiées pour trouver les thèmes émergents. Cependant, en procédant ainsi, ce système ne permet pas de trouver une information inattendue sur un site Web dès la première visite. Matsumura *et al.* [Matsumura *et al.*, 2001] ont également développé un système de recherche de thèmes émergents entre des communautés Web. Après avoir construit par classification des communautés composées de membres ayant les mêmes centres d'intérêt, le système analyse et visualise les co-citations entre des pages Web à l'aide de l'algorithme *KeyGraph* [Ohsawa *et al.*, 1998]. Les thèmes émergents correspondent alors aux pages Web intéressant plusieurs communautés. La faiblesse de ce système est qu'il suppose que de telles communautés puissent être définies et que les pages considérées puissent leur être attribuées.

Plus récemment, des travaux ont porté sur la détection d'information nouvelle, notamment dans le cadre de la compétition *TREC* [2]. Mais, le corpus proposé pour *TREC 2003* est composé de phrases et non de textes. De plus, la liste des documents fournis est triée par ordre chronologique. Ainsi, le problème posé revient plutôt à chercher des nouveaux documents dans le temps. Un certain nombre de systèmes proposés

1. http://www.nist.gov/speech/tests/tdt

2. Le challenge "novelty detection" est apparu pour la première fois lors de la conférence TREC 2002. Les communications présentées à cette conférence et aux suivantes sont disponibles à l'adresse http://trec.nist.gov

[Soboroff et Harman, 2003] peuvent alors identifier une phrase pertinente en la comparant, au moyen d'un critère de similarité, à celles qui la précèdent et à celles qui la suivent dans le corpus ; ce qui n'est pas réalisable pour le traitement de banques de textes intégraux scientifiques.

WebCompare développé par Liu *et al.* [Liu *et al.*, 2001] est probablement le système se rapprochant le plus de nos travaux. Il s'agit d'un système destiné à la veille concurrentielle. Après que l'utilisateur ait indiqué les adresses (*URL*) de pages Web de ses concurrents, *WebCompare* est capable de trouver les pages contenant des informations inattendues par rapport à celles figurant sur son propre site. Le caractère inattendu d'une page Web est évalué à l'aide d'une mesure basée sur le paradigme TF.IDF, tout comme les mesures que nous proposons dans cet article.

L'architecture globale du système automatique de veille technologique que nous avons développé est décrite dans la section suivante, tandis que les différentes mesures d'extraction d'information inattendue à partir de textes sont définies dans la troisième section. Nous présentons dans la quatrième section les résultats des expérimentations conduites avec chacune de ces mesures sur une base de résumés d'articles scientifiques, avant de conclure dans la dernière section, sur les perspectives ouvertes par cette recherche.

2 Le système UnexpectedMiner

Dans le cadre de la veille technologique, nous avons développé le système *UnexpectedMiner* qui vise à extraire, de corpus documentaires, des documents pertinents pour le veilleur en ce sens qu'ils traitent de sujets jusque là inattendus et inconnus de celui-ci et, plus largement, de la communauté spécialisée dans le domaine considéré. De plus, le système doit prendre en compte explicitement la demande du veilleur tout en ne lui imposant pas une forte participation. Finalement, un aspect important que nous avons souhaité conférer à notre système est qu'il ne soit pas dédié à un domaine ou à un sujet particulier.

Compte tenu de ces objectifs, nous proposons un système articulé autour de plusieurs modules, représenté par la figure 1.

2.1 Pré-traitement des données

Dans un premier temps, le responsable de la cellule de veille spécifie ses besoins en proposant quelques documents de référence. Dans la suite de cet article, l'ensemble de ces documents sera noté R et $|R|$ désignera leur nombre. Dans la pratique, entre dix et vingt documents doivent suffire pour cibler le domaine de la veille. Le système doit ensuite consulter des nouveaux documents dans divers corpus à sa disposition afin d'y rechercher les informations inattendues. Dans la suite N désignera l'ensemble de ces nouveaux documents et $|N|$ son cardinal. Les ensembles R et N doivent ensuite subir un pré-traitement. Le module conçu à cet effet comporte un certain nombre de traitements classiques tels qu'un nettoyage pour éliminer les éléments non pertinents des documents (logo, url, balises, ...), une analyse morphologique des mots des phrases

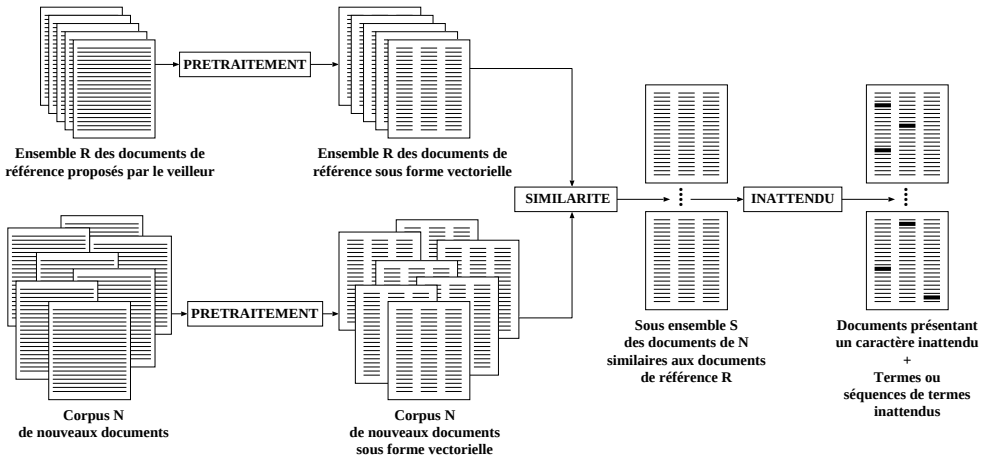

FIG. 1 – *Architecture du système UnexpectedMiner*

extraites et la suppression des mots vides. Finalement, chaque document est représenté classiquement sous forme vectorielle. Le document d_j est ainsi considéré comme un ensemble de mots indexés t_i où chaque mot indexé est en fait un mot du document d_j. Un index référence les m mots t_1,t_2,...,t_m rencontrés dans l'ensemble des documents.

Chaque document est représenté par un vecteur de poids $\vec{d_j} = (w_{1,j},w_{2,j},...,w_{m,j})$ où $w_{i,j}$ représente le poids du mot t_i dans le document d_j. Si le mot t_i n'apparaît pas dans le document d_j alors $w_{i,j} = 0$. Pour évaluer le poids d'un mot dans un document nous utilisons la formule classique TF.IDF [Salton et McGill, 1983].

TF (Term Frequency) correspond à la fréquence relative du mot t_i dans un document d_j définie par :

$$tf_{i,j} = \frac{f_{i,j}}{max_l f_{l,j}}$$

où $f_{i,j}$ désigne la fréquence du mot t_i dans le document d_j. Plus le mot t_i est fréquent dans le document d_j, plus $tf_{i,j}$ est élevé.

IDF (Inverse Document Frequency) est une mesure du pouvoir discriminant du mot t_i définie par :

$$idf_i = \log_2 \frac{Nd}{n_i} + 1$$

où Nd est le nombre de documents traités et n_i le nombre de documents contenant le mot t_i. Plus le mot t_i est rare dans l'ensemble des documents, plus idf_i est élevé. Dans la pratique, IDF est calculée simplement par :

$$idf_i = \log \frac{Nd}{n_i}$$

Le poids $w_{i,j}$ d'un mot t_i dans un document d_j est alors obtenu en combinant les deux critères précédents :

$$w_{i,j} = tf_{i,j} \times idf_i$$

Ce poids est d'autant plus élevé que le mot t_i est fréquent dans le document d_j et rare dans les autres documents.

2.2 Recherche de documents similaires

Le but du second module est d'extraire de la base N de nouveaux documents, ceux qui sont le plus similaires aux documents de référence R fournis par le veilleur. La similarité s_{jk} entre un nouveau document $d_j \in N$ et un document de référence $d_k \in R$ est égale à la distance du *cosinus*, couramment employée dans les systèmes de recherche d'information. Elle est égale au cosinus de l'angle formé par les vecteurs représentant ces documents :

$$s_{jk} = \frac{\vec{d_j} \bullet \vec{d_k}}{|\vec{j}| \times |\vec{k}|}$$

où

$$\vec{d_j} \bullet \vec{d_k} = \sum_i w_{i,j} \times w_{i,k}$$

$$|\vec{j}| = \sqrt{\sum_{i=1,m} w_{i,j}^2}$$

Comme rien ne garantit qu'un nouveau document $d_j \in N$, très similaire à un des documents de référence, le sera aussi des autres, nous calculons la similarité moyenne s_j du nouveau document $d_j \in N$ avec l'ensemble des documents de référence R par :

$$s_j = \frac{1}{|R|} \sum_{k=1}^{|R|} s_{jk}$$

Ainsi, s_j sera élevée si d_j est globalement proche des documents de référence tandis qu'elle sera faible si d_j est seulement proche de quelques uns.

Après avoir classé par ordre décroissant de similarité moyenne les nouveaux documents, un sous ensemble S est extrait de N. Il est composé des nouveaux documents les plus proches de ceux fournis comme référence par le veilleur. Le nombre de documents similaires extraits est choisi au vu de la liste des titres des documents : il ne doit pas être trop faible pour ne pas écarter des documents contenant des informations inattendues, ni trop élevé pour ne pas conserver des documents hors sujet. Il pourrait aussi être fixé par un apprentissage préalable réalisé sur un échantillon du corpus.

2.3 Recherche d'information inattendue

Le module de recherche d'information inattendue constitue le coeur du système *UnexpectedMiner* et la partie la plus originale de ce travail. L'objectif de ce module est, en effet, de rechercher les documents de S contenant des informations inattendues par rapport à celles contenues non seulement dans les documents de référence (R) mais

aussi dans les documents de S sélectionnés à l'étape précédente. Or, un document sera très inattendu si les thèmes qu'il aborde ne sont présents ni dans un autre document de S ni dans un document de R. Pour identifier de telles informations, ce module repose sur de nouvelles mesures, décrites dans la section suivante, et qui peuvent être rapprochées d'autres travaux qui se sont intéressés à la même problématique tels que ceux de [Cherfi et Toussaint, 2002, Cherfi *et al.*, 2003] ou [Azé, 2003].

3 Mesures du caractère inattendu d'un document

Cinq mesures ont été proposées pour évaluer le caractère inattendu d'un document.

3.1 Mesure de Liu : M1

La première mesure est directement inspirée du critère proposé par Liu, Ma et Yu [Liu *et al.*, 2001] pour repérer des pages inattendues dans un site WEB. Elle est définie par :

$$M1(d_j) = \frac{\sum_{i=1}^{m} U^1_{i,j,c}}{m}$$

avec:

$$U^1_{i,j,c} = \begin{cases} 1 - \frac{tf_{i,c}}{tf_{i,j}} & si\ tf_{i,c}/tf_{i,j} \leq 1 \\ 0 & sinon \end{cases}$$

où d_j désigne un document de S et D_c le document obtenu en combinant tous les documents de référence de R avec les documents sélectionnés sauf d_j : $R \cup S - \{d_j\}$.

L'inconvénient de la mesure de Liu ($M1$) est qu'elle prend la même valeur pour deux mots t_i et $t_{i'}$ apparaissant avec des fréquences différentes dans un nouveau document $d_j \in S$ dès lors que ces mots n'apparaissent pas dans D_c (autrement dit dans les autres documents de $R \cup S - \{d_j\}$). Or, il serait souhaitable d'obtenir une valeur d'inattendu $U^1_{i,j,c}$ pour t_i supérieure à $U^1_{i',j,c}$ calculée pour $t_{i'}$ si t_i est plus fréquent que $t_{i'}$ dans d_j, notamment dans le cas où t_i correspond à un mot encore jamais rencontré, car il traduit un nouveau concept, alors que $t_{i'}$ est un mot mal orthographié. Cette remarque nous a conduit à proposer et à expérimenter d'autres mesures pour évaluer le caractère inattendu d'un document.

3.2 Mesure tenant compte de la fréquence des mots : M2

Dans cette seconde mesure, le caractère inattendu d'un mot t_i dans un document $d_j \in S$ par rapport à l'ensemble des autres documents D_c est définie par :

$$U^2_{i,j,c} = \begin{cases} tf_{i,j} - tf_{i,c} & si\ tf_{i,j} - tf_{i,c} \geq 0 \\ 0 & sinon \end{cases}$$

Le caractère inattendu d'un document d_j est, comme dans la mesure de Liu ($M1$), égal à la moyenne des mesures d'inattendu associées aux mots représentant d_j :

$$M2(d_j) = \frac{\sum_{i=1}^{m} U^2_{i,j,c}}{m}$$

Cette seconde mesure comble la lacune de la première. En effet, en reprenant l'exemple précédent, si le mot t_i figure plus fréquemment que $t_{i'}$ dans le document d_j sans que ni l'un ni l'autre n'apparaissent dans D_c alors :

$$U^2_{i,j,c} > U^2_{i',j,c}$$

3.3 Mesure tenant compte de la fréquence des termes : M3

Dans la mesure précédente, les mots ont été considérés individuellement. Cependant, dans le domaine de la veille, comme en recherche d'information, c'est souvent l'association de plusieurs mots, telle que par exemple "data mining" qui est intéressante. Ceci nous a conduit à représenter chaque document par des mots et par des termes c'est à dire par des séquences de mots indissociables. Il existe un certain nombre d'outils, tels que ANA [Enguehard et Pantéra, 1995], ACABIT [Daille, 2003], FASTR [Jacquemin, 1999] qui permettent d'extraire des candidats termes, en lien avec des bases terminologiques. Mais, dans le contexte de la recherche de signaux faibles, notre but était plus particulièrement d'identifier des associations de mots inhabituelles, employées pour exprimer un nouveau concept et, n'ayant pas encore été fixées dans une terminologie. De plus, il s'agissait de limiter les traitements linguistiques pour permettre une exploitation de corpus de taille conséquente en temps raisonnable. C'est pourquoi, plutôt que d'employer un outil d'indexation automatique intégrant un corpus de termes prédéfini, nous avons utilisé une implémentation d'un algorithme d'extraction de motifs séquentiels fréquents, basé sur celui d'Agrawal et Srikant [Agrawal et Srikant, 1995], pour extraire des séquences fréquentes de mots dans les documents.

La prise en compte de telles séquences nous a amené à définir une troisième mesure, qui est une adaptation de la mesure tenant compte de la fréquence des mots (M2) dans laquelle :

$$tf_{i,j} = \frac{f_{i,j}}{max_l f'_{l,j}}$$

où $max_l f'_{l,j}$ est la fréquence maximale observée dans les mots et les séquences de mots.

On peut cependant observer que les mesures précédentes ne tiennent pas compte du pouvoir discriminant d'un mot exprimé par idf. Pour cette raison, il nous a paru intéressant de concevoir des mesures d'inattendu qui exploitent directement cette information. C'est le cas des deux mesures suivantes.

3.4 Mesure tenant compte du pouvoir discriminant des mots : M4

La quatrième mesure fait intervenir directement le pouvoir discriminant idf_i d'un mot t_i puisqu'elle évalue le caractère inattendu d'un document d_j par la somme des poids $w_{i,j}$ des mots t_i qui le représentent :

$$M4(d_j) = \sum_{i=1}^{m} w_{i,j}$$

Toutefois, avec cette mesure deux documents d_j et d'_j peuvent présenter la même valeur d'inattendu alors que les poids des mots représentatifs du premier document sont égaux tandis que ceux du second document sont très différents.

3.5 Mesure tenant compte du poids maximum : M5

Pour pallier à la limite de la mesure tenant compte du pouvoir discriminant des mots ($M4$), la cinquième mesure proposée attribue comme valeur d'inattendu à un document d_j le poids le plus élevé apparu dans son vecteur de représentation :

$$M5(d_j) = max_l w_{l,j}$$

Des expérimentations ont été réalisés pour évaluer notre système et comparer entre elles ces différentes mesures. Elles sont présentées dans la section suivante.

4 Expérimentations

4.1 Corpus et critères d'évaluation utilisés

L'ensemble de référence R est composé de 18 articles scientifiques en anglais consacrés à l'apprentissage automatique (Machine learning) mais dont aucun n'abordent certains thèmes tels que *Support Vector Machines*, *Affective Computing*, *Reinforcement Learning*, etc. La base N est composée de 57 nouveaux documents dont 17 sont considérés par le veilleur comme similaires aux documents de référence. Parmi ces 17 documents, 14 traitent de thèmes jugés inattendus par ce dernier.

Pour évaluer *UnexpectedMiner* nous avons utilisé les critères de *précision* et de *rappel* définis par Swets [Swets, 1963].

Documents	Inattendus	Non Inattendus
Extraits	n_1	n_3
Non extraits	n_2	n_4

La **Précision** mesure le pourcentage d'information extraite qui est correcte :

$$Précision = \frac{n_1}{n_1 + n_3}$$

Dans le module d'extraction d'information inattendue de notre système, elle donne le pourcentage de documents qui ont réellement un caractère inattendu parmi les documents extraits par le système.

Le **Rappel** mesure le pourcentage d'information recherchée correctement extraite par le système :

$$Rappel = \frac{n_1}{n_1 + n_2}$$

Dans le module d'extraction d'information inattendue de notre système, il évalue la capacité du système à retrouver les documents contenant des informations inattendues

dans le corpus S.

De plus, la précision et le rappel sont calculés d'abord en demandant au système d'extraire le document le plus inattendu, puis en lui demandant d'extraire les deux documents les plus inattendus et ainsi de suite. De la sorte, en considérant un ensemble de documents de plus en plus grand, il est possible de tracer des courbes de précision et de rappel où l'axe des abscisses indique le nombre de documents demandés au système.

4.2 Evaluation des cinq mesures

L'apport principal de ce travail étant la définition de nouvelles mesures du caractère inattendu d'un document, le module qui met en oeuvre ces mesures a d'abord été évalué indépendamment du module d'extraction de documents similaires.

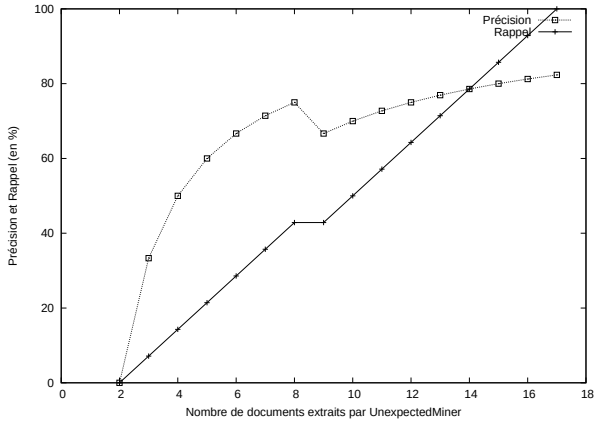

FIG. 2 – *Précision et Rappel pour la mesure 1*

Pour ce faire, nous avons dans un premier temps restreint la base S aux 17 nouveaux documents jugés similaires, par le veilleur, aux documents de référence R. Les résultats obtenus en termes de rappel et de précision à l'aide des cinq mesures définies précédemment sont présentés dans les figures 2 à 6 où l'axe des abscisses indique le nombre de documents extraits par le système.

Alors que la base N comporte très majoritairement des documents qui traitent de sujets inattendus (14 documents sur 17), seule la mesure de Liu ($M1$) ne parvient pas à les retrouver en priorité puisque la précision vaut 0% en ne considérant que les deux premiers documents extraits (figure 2) alors qu'elle atteint 100% pour les autres mesures (figures 3 à 6).

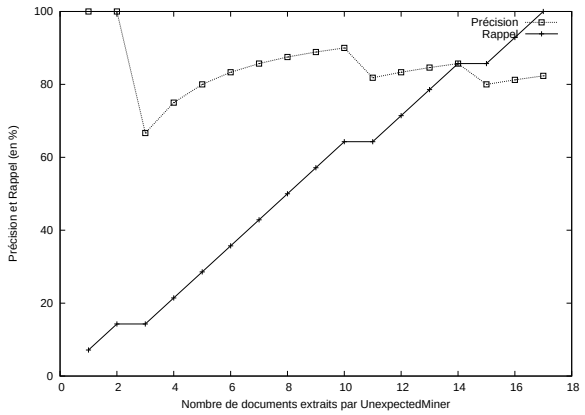

FIG. 3 – *Précision et Rappel pour la mesure 2*

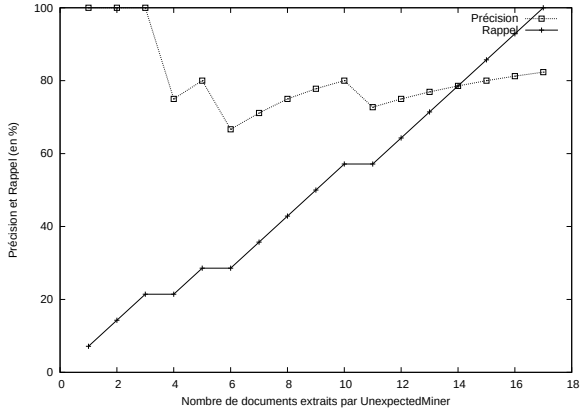

FIG. 4 – *Précision et Rappel pour la mesure 3*

Les résultats obtenus à l'aide des mesures tenant compte de la fréquence des mots ($M2$: figure 3) et des termes ($M3$: figure 4) sont plus satisfaisants. Ce sont toutefois les mesures tenant compte du pouvoir discriminant ($M4$) et du poids maximum ($M5$) qui fournissent en priorité le plus grand nombre de documents traitant de sujets inattendus.

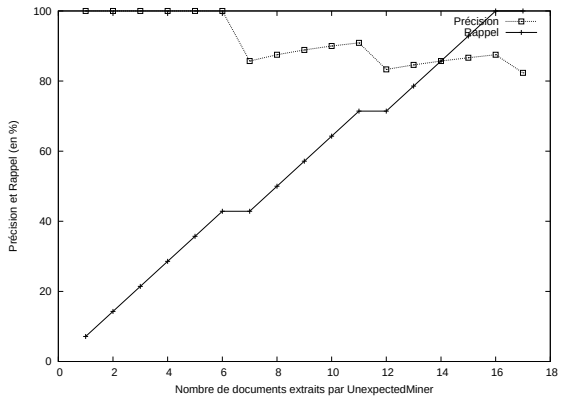

FIG. 5 – *Précision et Rappel pour la mesure 4*

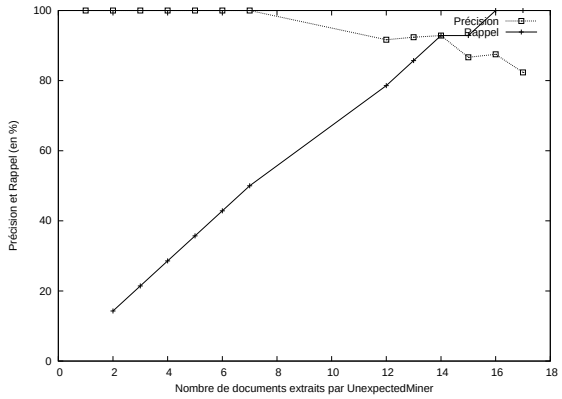

FIG. 6 – *Précision et Rappel pour la mesure 5*

La précision reste en effet égale à 100% lorsqu'on considère jusqu'à six documents pour la mesure tenant compte du pouvoir discriminant ($M4$: figure 5) et jusqu'à sept pour la mesure du poinds maximum ($M5$: figure 6).

4.3 Evaluation globale du système

Nous avons ensuite considéré le système complet en procédant non seulement à la recherche de documents inattendus mais aussi à l'extraction des documents similaires (module 2 et 3). L'évaluation porte alors sur la base N contenant les 57 nouveaux documents dont 17 sont considérés par le veilleur comme similaires aux documents de référence. Parmi ces 17 rappelons que 14 traitent de thèmes jugés inattendus par l'expert. Lors de la phase d'extraction de documents similaires, parmi les 15 premiers documents jugés similaires aux documents de référence par le système, 9 seulement

l'étaient réellement ; ce qui correspond à un taux de précision de 60 % et à un taux de rappel de 52,9 %. Ainsi, les erreurs commises par le système dans le module 2 constitue du bruit rendant plus difficile la tâche d'extraction d'information inattendue. Il s'agit en effet de documents qui sont hors sujet et qui ont de fortes chances d'être considérés ensuite à tort par le système comme étant inattendus. Parmi les 9 documents effectivement similaires identifiés pat le système, 7 abordaient des thèmes inattendus.

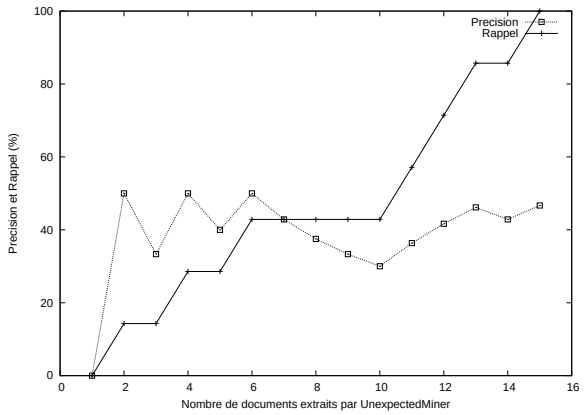

FIG. 7 – *Evaluation globale du système - Précision et Rappel pour la mesure 1*

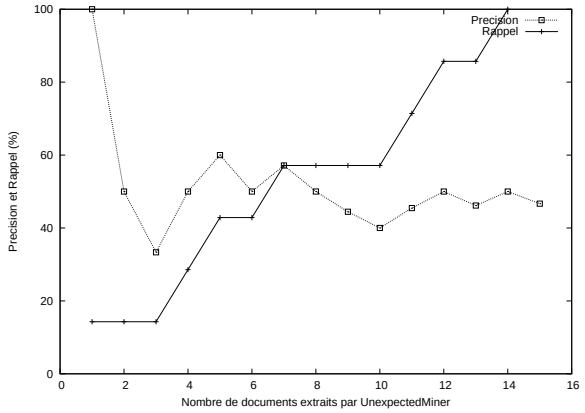

FIG. 8 – *Evaluation globale du système - Précision et Rappel pour la mesure 2*

Dans cette expérimentation, comme précédemment, seule la mesure de Liu ($M1$: figure 7) n'est pas capable d'extraire en premier un document traitant un sujet inattendu : la précision est égale à 0% alors qu'elle vaut 100% pour les autres mesures ($M2$, $M3$, $M4$ et $M5$ respectivement figures 8, 9, 10, 11) qui identifient correctement le même document inattendu. Notons que la mesure de Liu ($M1$) détecte moins bien

les documents inattendus puisque le rappel atteint 100% uniquement lorsque le nombre de documents extraits devient égal au nombre de documents fournis au système.

Les performances de la mesure tenant compte de la fréquence des mots ($M2$: figure 8) et de la mesure tenant compte du pouvoir discriminant ($M4$: figure 10) sont assez comparables mais c'est encore la mesure du poids maximum ($M5$: figure 11) qui extrait en priorité les documents se rapportant à des sujets inattendus. En revanche cette mesure présente la particularité d'attribuer relativement souvent une même valeur à plusieurs documents.

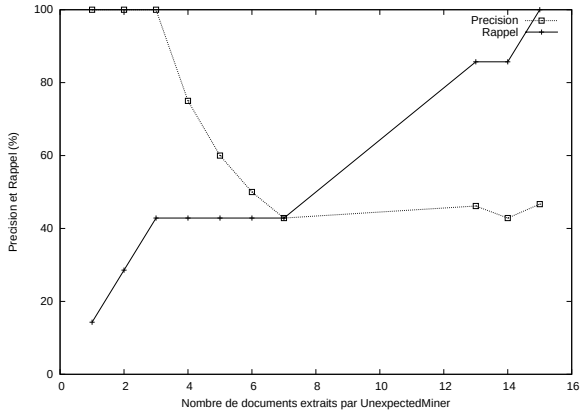

FIG. 9 – *Evaluation globale du système - Précision et Rappel pour la mesure 3*

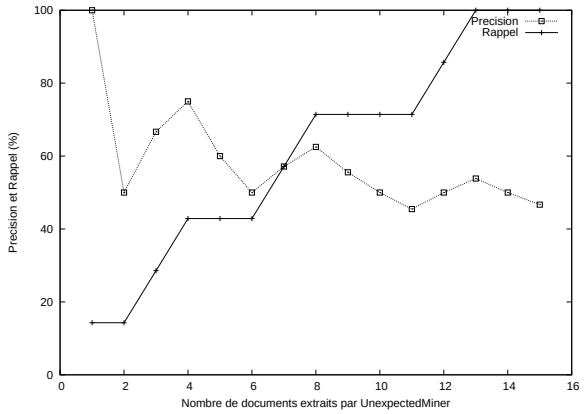

FIG. 10 – *Evaluation globale du système - Précision et Rappel pour la mesure 4*

Enfin, si les résultats fournis en prenant en compte les termes ($M3$: figure 9) sont un peu moins satisfaisants, par contre, les séquences de mots inattendues retrouvées correspondent bien à celles recherchées à savoir "support vector machine" ou "reinforcement

learning". A ce propos, il convient de noter que le système *UnexpectedMiner* présente l'avantage d'indiquer les mots ou les séquences de mots qui ont le plus contribué à faire d'un document qui lui est soumis un document inattendu.

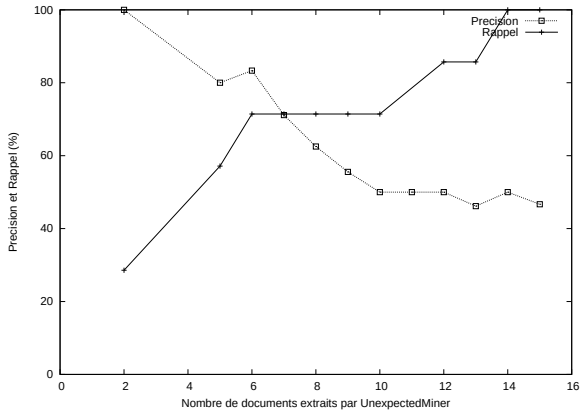

FIG. 11 – *Evaluation globale du système - Précision et Rappel pour la mesure 5*

5 Conclusion et perspectives

Nous avons développé un système de veille qui vise à extraire d'un corpus documentaire des documents pertinents dans le sens où ils traitent de sujets inattendus et inconnus du veilleur auparavant. Plusieurs mesures du caractère inattendu d'un document ont été proposées et comparées. Bien que les résultats obtenus soient encourageants, ils sont encore loin d'être totalement satisfaisants. Ces expérimentations ont toutefois permis d'envisager plusieurs améliorations du système.

D'une part, nous avons pu constater combien il est difficile, pour le système, de repérer des documents contenant des informations inattendues dans un ensemble comportant des documents jugés à tort comme similaires par le système. Pour trouver des informations intéressantes pour le veilleur, il paraît donc indispensable de cibler correctement l'ensemble des documents dans lequel il faut rechercher ces informations. Ceci conduit à accorder une attention particulière aux premiers modules consacrés à la représentation des documents et à l'extraction de l'ensemble S des documents jugés similaires, par le système, aux documents de référence fournis par le veilleur. Il serait intéressant d'étudier d'autres mesures de similarité [Lebart et Rajman, 2000] ou encore un modèle probabiliste de représentation des documents [Siolas et D'Alché-Buc, 2003].

D'autre part, une autre amélioration du système pourrait être liée à la prise en compte de la structure des documents [Piwowarski *et al.*, 2002]. Dans le contexte de la veille stratégique, ceci paraît d'autant plus facile à intégrer que la plupart des bases utilisées contiennent des documents fortement structurés. Par exemple, les articles scientifiques ou les résumés de thèses sont composés de parties clairement identifiées comme le titre, les auteurs, la liste des mots clés, le résumé et le document lui même,

généralement organisés en sections subdivisées en paragraphes. Il en va de même pour les fiches descriptives de brevets. Dans le cas de documents diffusés sur le Web, qui sont de plus en plus largement exploités en veille, ceci s'avère également vrai du fait de l'utilisation de langages de description tels que XML qui permettent de représenter conjointement l'information textuelle et l'information sur la structure du document. Dans le contexte d'une veille, le changement de position d'un mot dans les différentes parties des documents au cours du temps pourrait être un critère pertinent pour déceler automatiquement des changements de tendance ou des évolutions du sujet.

Références

[Agrawal et Srikant, 1995] R. Agrawal et R. Srikant. Mining sequential patterns. In *Eleventh International Conference on Data Engineering*, pages 3–14, Taipei, Taiwan, 1995. IEEE.

[Azé, 2003] J. Azé. Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances. *Extraction des connaissances et apprentissage, Hermès*, 17(1):171–182, 2003.

[Bun et Ishizuka, 2001] K. K. Bun et M. Ishizuka. Emerging topic tracking system. In *Proceedings of the International Conference on Web Intelligence*, LNAI 2198, pages 125–130, 2001.

[Cherfi *et al.*, 2003] H. Cherfi, A. Napoli, et Y. Toussaint. Vers une méthodologie de fouille de textes s'appuyant sur l'extraction de motifs fréquents et de règles d'association. In *Actes de la Conférence d'Apprentissage Automatique (CAP 2003)*, pages 61–76, 2003.

[Cherfi et Toussaint, 2002] H. Cherfi et Y Toussaint. Fouille de textes par combinaison de règles d'association et d'indices statistiques. In *Actes du Premier Colloque International sur la Fouille de texte CIFT*, pages 67–80, 2002.

[Daille, 2003] B. Daille. Conceptual structuring through term variations. In *Proceedings of ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 9–16, September 2003.

[Desvals et Dou, 1992] H. Desvals et H. Dou. *La veille technologique.* Dunod, 1992.

[Enguehard et Pantéra, 1995] C. Enguehard et L. Pantéra. Automatic natural acquisition of a terminology. *Journal of quantitative linguistics*, 2(1):27–32, 1995.

[Fayyad *et al.*, 1996] U.M Fayyad, G. Piatetsky, P. Smyth, et R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining.* AAAI/MIT Press, 1996.

[Feldman et Dagan, 1995] R. Feldman et Ido Dagan. Knowledge discovery from textual databases. In *Proceedings of the International Conference on Knowledge Discovery from DataBases*, pages 112–117, 1995.

[François, 2003] C. François. Outils de veille : typologie. In *Rencontres des professionnels de l'IST (Paris)*, 2003.

[Jacquemin, 1999] C. Jacquemin. Syntagmatic and paradigmatic representations of term variation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 341–348, 1999.

[Jakobiak, 1990] F. Jakobiak. *Pratique de la veille technologique.* Editions d'Organisation, 1990.

[Jakobiak, 1994] F. Jakobiak. *Le brevet source d'information.* Dunod, 1994.

[Kodratoff, 1999] Y. Kodratoff. Knowledge discovey in texts: A definition and applications. In *Proceedings of the International Symposium on Methodologies for Intelligent Systems*, volume LNAI 1609, pages 16–29, 1999.

[Lebart et Rajman, 2000] L. Lebart et M. Rajman. Computing similarity. In *Handbook of Natural Language Processing*, pages 477–505. Dekker, 2000.

[Lent *et al.*, 1997] B. Lent, R. Agrawal, et R. Srikant. Discovering trends in text databases. In *Proceedings KDD'97*, pages 227–230. AAAI Press, 14–17 1997.

[Liu *et al.*, 2001] B. Liu, Y. Ma, et P. S. Yu. Discovering unexpected information from your competitors' web sites. In *Proceedings KDD'2001*, pages 144–153, 2001.

[Losiewicz *et al.*, 2000] P. Losiewicz, D.W. Oard, et R. Kostoff. Textual data mining to support science and technology management. *Journal of Int. Inf. Systems*, 15:99–119, 2000.

[Matsumura *et al.*, 2001] N. Matsumura, Y. Ohsawa, et M. Ishizuka. Discovery of emerging topics between communities on WWW. In *Proceedings Web Intelligence'2001*, pages 473–482, Maebashi, Japan, 2001. LNCS 2198.

[Ohsawa *et al.*, 1998] Y. Ohsawa, N. E. Benson, et M. Yachida. Keygraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In *Proceedings of the Advances in Digital Libraries Conference*, pages 12–18, 1998.

[Piwowarski *et al.*, 2002] B. Piwowarski, L. Denoyer, et P. Gallinari. Un modèle pour la recherche d'information sur les documents structurés. In *6ème Journées internationales d'Analyse statistique de Données Textuelles*, 2002.

[Rajaraman et Tan, 2001] K. Rajaraman et A.H. Tan. Topic detection, tracking and trend analysis using self-organizing neural networks. In *Proceedings PAKDD'2001*, pages 102–107, Hong-Kong, 2001.

[Salton et McGill, 1983] G. Salton et M. J. McGill. Introduction to modern information retrieval. In *McGraw-Hill*, 1983.

[Sebastiani, 2002] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, March 2002.

[Siolas et D'Alché-Buc, 2003] G. Siolas et F. D'Alché-Buc. Modèles probabilistes et scores de Fisher pour la représentation de mots et de documents. In *Actes de la Conférence d'Apprentissage CAP 2003*, pages 47–59, 2003.

[Soboroff et Harman, 2003] I. Soboroff et D. Harman. Overview of the trec 2003 novelty track. In *NIST Special Publication:SP 500-255*, pages 38–53. The Twelfth Text Retrieval Conference (TREC 2003), 2003.

[Swets, 1963] J.A. Swets. Information retrieval systems. *Science*, 141:245–250, 1963.

[Zhu *et al.*, 1999] D. Zhu, A.L. Porter, S. Cunningham, J. Carlisie, et A. Nayak. A process for mining science and technology documents databases, illustrated for the case of "knowledge discovery and data mining". *Ciencia da Informação*, 28(1):7–14, 1999.

[Zhu et Porter, 2002] D. Zhu et A.L. Porter. Automated extraction and visualization of information for technological intelligence and forecasting. *Technological Forecasting and Social Change*, 69:495–506, 2002.

Summary

In the domain of business intelligence, computers are useful for extracting scientific or technological information that may be relevant to companies. Moreover, in this context, the aim is to find some unexpected knowledge that may appear with a low frequency. In order to automatically discover some useful knowledge from databases (patents, research publications,etc) we propose to use text mining techniques. Nevertheless, most of these techniques can help finding some frequent information instead of unexpected one, thus, they are not well suited for business intelligence that requires a specific approach. To this end, we have designed several new knowledge discovery measures and integrated them in the UnexpectedMiner System that is able to extract some novel information that may be of interest for the user. We have experimented UnexpectedMiner on a database of scientific abstracts and reported the impact of the various measures on the efficiency of the system.