Extraction de connaissance inattendue Application à la veille technologique

François Jacquenet(*), Christine Largeron(**)

Université Jean Monnet de Saint-Etienne
(*)EURISE
23 rue du docteur Paul Michelon
(**)CREUSET
6, rue Basse des Rives
42023 Saint-Etienne Cedex 2
France
Francois.Jacquenet@univ-st-etienne.fr
Christine.Largeron@univ-st-etienne.fr

Résumé. Le domaine de la veille technologique vise à récolter, traiter, et analyser des informations scientifiques et techniques utiles aux acteurs économiques. Dans cet article, nous proposons d'utiliser des techniques de fouille de textes pour automatiser le processus de traitement des données issues de bases de textes scientifiques. Toutefois, la veille introduit une difficulté inhabituelle par rapport aux domaines d'application classiques des techniques de fouille de textes, puisqu'au lieu de rechercher de la connaissance fréquente cachée dans les données, il faut rechercher de la connaissance inattendue. 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 dans le domaine de l'apprentissage automatique.

Mot-clés : Fouille de textes, Extraction de connaissances à partir de données, Veille scientifique et technique

Keywords : Text mining, Knowledge extraction in databases, Science and technology watch

Palabra clave : Explotación del texto, Extracción de información a partir de los datos, Vigilencia scientifica y tecnologica

Thème retenu : Fouille d'information / data mining

1 Introduction

Depuis quelques années le secteur économique a pris conscience des enjeux liés à la maîtrise de l'information stratégique. Toutefois, les entreprises sont de plus en plus submergées d'informations. Elles ont de grandes difficultés à dégager les données stratégiques dont elles ont besoin pour anticiper les marchés, prendre des décisions et agir sur leur environnement socio-économique [Samier et Samoval, 2001, Martinet et Marti, 2001, Revelli, 2000, Carayon, 2003]. Ceci a conduit à l'émergence de l'intelligence économique définie par H. Martre comme "l'ensemble des actions de recherche, de traitement, de distribution et de protection de l'information obtenue légalement et utile aux acteurs économiques" [Martre, 1994]. Lorsque les informations à analyser sont de nature scientifique et technique, on parle plus spécifiquement de veille technologique pour désigner la surveillance des brevets et de la documentation scientifique (articles, thèses, ...) [Desvals et Dou, 1992, Jakobiak, 1990, Jakobiak, 1994].

Le processus de veille peut être décomposé en quatre phases principales : l'audit des besoins, la collecte des données, le traitement des données et la synthèse et la diffusion des résultats. Dans cet article, nous nous intéressons principalement à la troisième phase. Pour automatiser le traitement des données collectées, les techniques de fouille de données semblent attractives et d'autant plus adaptées que la plupart de ces données sont disponibles sous une forme numérique.

La fouille de données a connu un fort développement depuis le milieu des années 1990 du fait de la mise au point de nouveaux algorithmes performants permettant de traiter de gros volumes de données dans le domaine commercial [Fayyad *et al.*, 1996]. 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 Ronan 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. Les premiers travaux réalisés en fouille de textes ont consisté à appliquer les algorithmes développés pour la fouille de données sans tenir compte de la spécificité des données considérées à savoir leur caractère textuel. Ainsi, par exemple, B. Lent [Lent *et al.*, 1997] a montré comment il était possible d'utiliser les méthodes d'extraction de séquences fréquentes pour découvrir de nouvelles tendances dans une base de données de brevets chez IBM. Depuis, d'autres travaux ont vu le jour dans le domaine de la veille. On peut citer par exemple ceux de Liu [Liu *et al.*, 2001], Rajaraman [Rajaraman et Tan, 2001] ou encore Matsumura [Matsumura *et al.*, 2001] et dans le cadre français le projet Communication [Poibeau, 2003].

Toutefois, les algorithmes d'extraction de motifs séquentiels fréquents, employés habituellement en fouille de données, sont inappropriés pour effectuer de la veille surtout en raison même de la spécificité de ce domaine. Comme leur nom l'indique, ces outils s'intéressent en effet aux informations qui apparaissent fréquemment dans une base de données. Or, dans le domaine de l'intelligence économique, il est essentiel de détecter des informations nouvelles et inattendues pour le veilleur. De telles informations n'apparaissent donc pas en général avec une fréquence élevée. C'est vraisemblablement une des raisons principales pour laquelle les logiciels commercialisés répondent mal actuel-

lement à l'attente des veilleurs.

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 inattendus et inconnus auparavant de celui-ci. 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. Notre système peut être rapproché d'autres travaux qui se sont intéressés à la même problématique tels que ceux de [Liu *et al.*, 2001] ou [Cherfi *et al.*, 2003] ou [Azé, 2003].

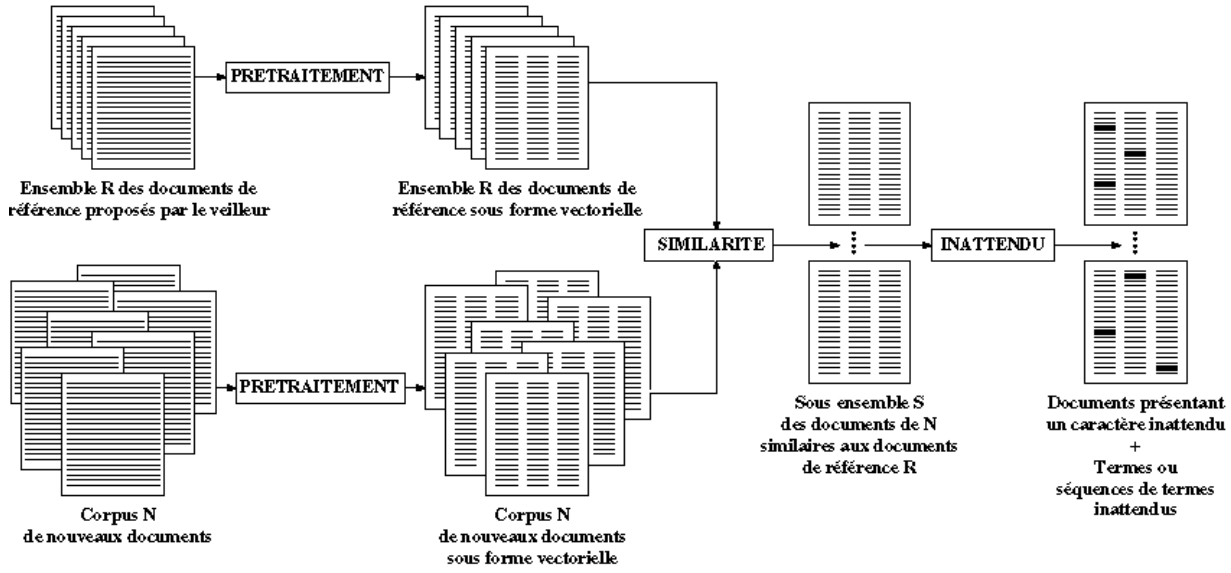

FIG. 1 – *Architecture du système UnexpectedMiner*

2.1 Pré-traitement des données

Dans un premier temps, le responsable de la cellule de veille spécifie ses besoins en produisant 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 innovantes. Dans la suite N désignera l'ensemble de ces nouveaux documents et $|N|$ son cardinal. Les ensembles R et N vont 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 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 termes indexés t_i où chaque terme indexé est en fait un mot du document d_j. Un index noté $T = t_1,t_2,...,t_m$ liste tous les termes rencontrés dans les documents. Chaque document est alors 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 terme t_i dans le document d_j. Si le terme t_i n'apparaît pas dans le document d_j alors $w_{i,j} = 0$. Pour évaluer le poids d'un terme dans un document la formule TF.IDF est généralement utilisée [Salton et McGill, 1983]. TF (Term Frequency) correspond à la fréquence relative du terme 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 terme t_i dans le document d_j. Plus le terme 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 terme t_i définie par :

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

où N est le nombre de documents traités et n_i le nombre de documents contenant le terme t_i. Plus le terme t_i est rare dans l'ensemble des documents, plus idf_i est élevé.

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

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

Il est d'autant plus élevé que le terme 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. La similarité moyenne s_j du nouveau document $d_j \in N$ avec l'ensemble des documents de référence R est égale à :

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

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.

2.3 Recherche d'information inattendue

Le module de recherche d'information inattendue constitue le coeur du système UnexpectedMiner. L'objectif de ce module est 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. En effet, 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. Ce module est décrit en détail dans la section suivante.

3 Mesures du caractère inattendu d'un document

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

3.1 Mesure 1

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 U^1 est qu'elle prend la même valeur pour deux termes t_i et $t_{i'}$ apparaissant avec des fréquences différentes dans un nouveau document $d_j \in S$ dès lors que ces termes 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 $U^1_{i,j,c}$ d'inattendu pour t_i supérieure à $U^1_{i',j,c}$ trouvée 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 nouveau mot clé 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 2

Dans cette seconde mesure, le caractère inattendu d'un terme 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 $M1$, égal à la moyenne des mesures d'inattendu associées aux termes représentant d_j :

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

Cette seconde mesure comble la lacune de la première. En effet, en reprenant l'exemple précédent, si le terme 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 :

$$U_{i,j,c}^2 > U_{i',j,c}^2$$

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

3.3 Mesure 3

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

$$M3(d_j) = \sum_{i=1}^{m} w_{i,j}$$

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

3.4 Mesure 4

Pour pallier la limite de $M3$, la quatrième mesure proposée attribue comme valeur d'inattendu à un document d_j le poids le plus élevé apparu dans son vecteur de représentation :

$$M4(d_j) = max_l w_{l,j}$$

3.5 Mesure 5

Dans le cas des mesures précédentes, seuls les termes ont été considérés. Or dans le domaine de la veille, comme en recherche d'information, c'est souvent l'association de plusieurs termes, telle que par exemple "data mining" qui est intéressante. Ceci nous a conduit à représenter chaque document par des termes et par des séquences de termes. Nous avons alors utilisé une extension de l'algorithme *apriori* – développé initialement par R. Agrawal et R. Srikant [Agrawal et Srikant, 1994] pour extraire des ensembles d'items fréquents servant à la construction de règles d'associations – proposée en 1995 [Agrawal et Srikant, 1995] pour extraire des ensembles de séquences fréquentes dans

les données. La prise en compte de ces séquences de termes nous a amené à définir une cinquième mesure, qui est une adaptation de 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 termes et les séquences de termes.

Des tests ont été réalisés pour évaluer ce système et comparer ces différentes mesures. Ils sont présentés 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, ... 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 J.A. Swets [Swets, 1963]. Dans le cas de notre système, la *précision* mesure le pourcentage de documents extraits par le système et qui ont réellement un caractère inattendu. Le *rappel* mesure quant à lui le pourcentage de documents ayant un caractère inattendu retrouvés dans le corpus N par le système. Ces critères sont classiques en recherche d'information et nous ne les détaillerons pas plus ici.

4.2 Protocole d'évaluation 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 puis globalement en tenant compte de tous les modules.

Dans un premier temps nous avons donc restreint la base S aux 17 nouveaux documents jugés similaires, par le veilleur, aux documents de référence de R. Les résultats obtenus en termes de rappel et de précision successivement à l'aide de chacune des cinq mesures définies précédemment, sont présentés dans les figures 2 à 6 où, l'axe des abscisses indique le nombre de documents demandés par l'utilisateur au système qui les restitue par valeur d'inattendu décroissante. Alors que la base N comporte très majoritairement des documents qui traitent de sujets inattendus (14 documents sur 17), seule la mesure $M1$ 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).

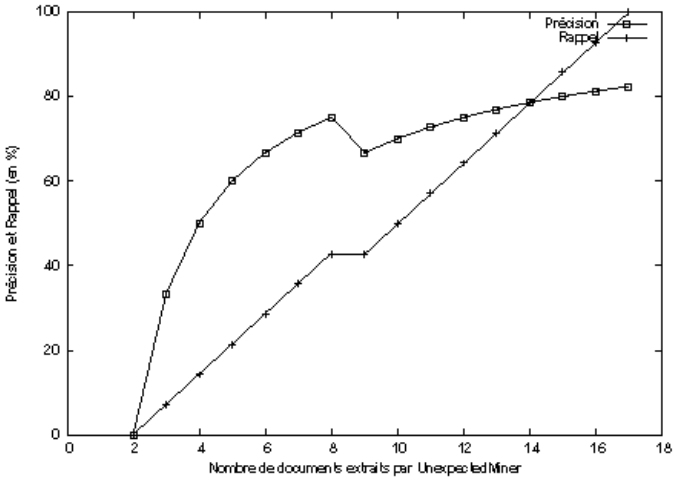

FIG. 2 – *Précision et Rappel pour la mesure 1*

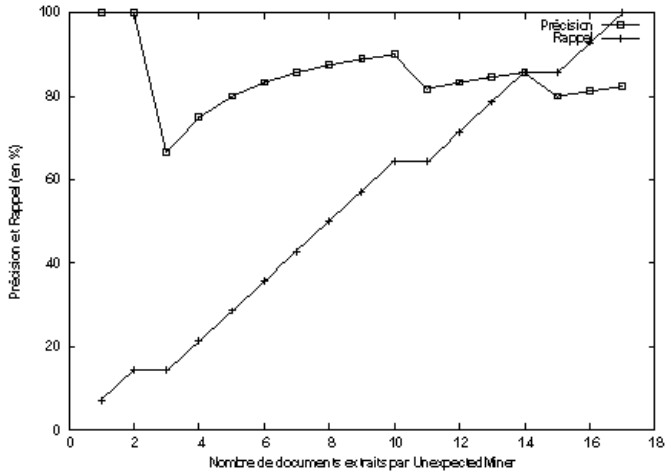

FIG. 3 – *Précision et Rappel pour la mesure 2*

Les résultats obtenus à l'aide des mesures $M2$ (figure 3) et $M5$ (figure 4) sont plus satisfaisants. Ce sont toutefois les mesures $M3$ et $M4$ qui fournissent en priorité le plus grand nombre de documents traitant de sujets inattendus. La précision reste en effet égale à 100% lorsqu'on considère jusqu'à six documents pour $M3$ (figure 5) et jusqu'à sept pour $M4$ (figure 6).

Nous avons ensuite considéré le système complet et l'avons évalué sur la base N contenant les 57 nouveaux documents. 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 %. Parmi ces 9 documents, 7 abordaient des thèmes inattendus. Dans cette seconde expérience encore, seule la mesure $M1$ 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 $M2$,

$M3$, $M4$ et $M5$ qui identifient correctement le même document inattendu. Notons que la mesure $M1$ 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 $M2$ et de $M3$ sont assez comparables mais c'est encore la mesure $M4$ 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. Enfin, si les résultats fournis par $M5$ 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 UnexpedtedMiner 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.

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. La première concerne le prétraitement qui, en plus de la lemmatisation et de l'élimination des mots vides, pourrait comporter une analyse linguistique plus poussée. La seconde porte sur la représentation des documents et l'extraction de documents jugés similaires, par le système, aux documents de référence fournis par le veilleur. Enfin, une autre voie d'amélioration pourrait être liée à la prise en compte de la structure des documents

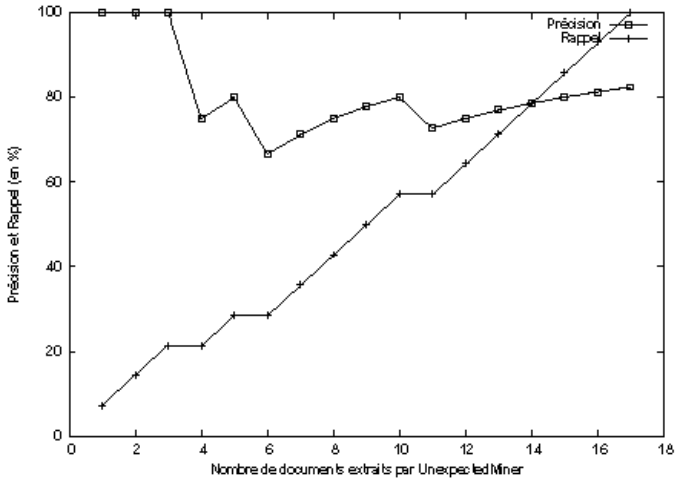

FIG. 4 – *Précision et Rappel pour la mesure 5*

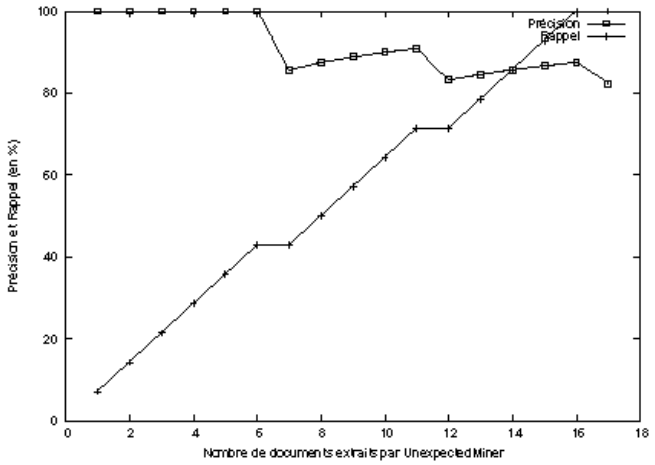

FIG. 5 – *Précision et Rappel pour la mesure 3*

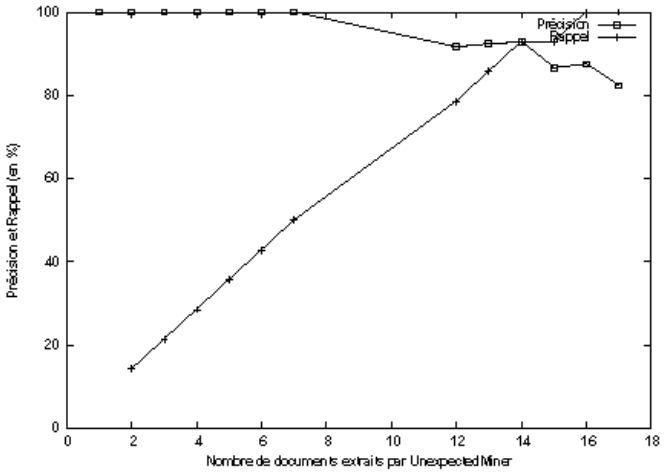

FIG. 6 – *Précision et Rappel pour la mesure 4*

Références

[Agrawal et Srikant, 1994] R. Agrawal et R. Srikant. Fast algorithms for mining association rules. In *Proceedings VLDB'94*, pages 487–499. Morgan Kaufmann, 1994.

[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.

[Carayon, 2003] B. Carayon. *Intelligence économique, compétitivité et cohésion sociale*. La documentation française, Paris, 2003.

[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.

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

[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.

[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.

[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.

[Martinet et Marti, 2001] B. Martinet et Y.M. Marti. *L'intelligence économique*. Editions de l'organisation, 2001.

[Martre, 1994] H. Martre. *Intelligence économique et stratégie des entreprises*. Commissariat Général au Plan. Rapport du groupe présidé par Henri Martre. La Documentation française, 1994.

[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.

[Poibeau, 2003] T. Poibeau. *Extraction automatique d'information. Du texte brut au web sémantique*. Hermès, 2003.

[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.

[Revelli, 2000] C. Revelli. *Intelligence Stratégique sur Internet*. Dunod, 2000.

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

[Samier et Samoval, 2001] H. Samier et V. Samoval. *La veille stratégique sur Internet*. Hermès, 2001.

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

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