ANALYSE DES DOCUMENTS TEXTE POUR LA SURVEILLANCE ET LA VEILLE TECHNOLOGIQUE EN TELECOMMUNICATION

Hamid MACHHOUR (*), Ismail KASSOU (*), Khalid EL HIMDI (**), Ilham BERRADA (*)
hamid.machhour@gmail.com, kassou@ensias.ma, elhimdi@menara.ma, iberrada@ensias.ma

(*) ENSIAS, Université Mohammed V Souissi, BP 713, Agdal, Rabat, Maroc
(**) Faculté des Sciences, Université Mohammed V Agdal, BP 1014 RP, Rabat, Maroc

Mots clefs :
Fouille de texte, veille technologique, traitement automatique des langues, outil de veille, ontologie de concepts.
Keywords:
Text mining, technical observation, language processing, monitoring tool, ontology concepts.
Palabras clave :
Minería de texto, Escudriñar tecnológico, procesamiento del lenguaje, herramienta de monitoreo, ontología concepto.

Résumé

Cet article montre une approche pratique de surveillance automatique de site web pour la veille technologique en télécommunication. Sa mise en ouvre consiste en la sélection de sources web à surveiller, un outil a été développé afin de collecter en local les documents pertinents. Puis, le corpus de document est analysé afin de détecter de nouvelles connaissances et de dégager le sens stratégique des informations collectées. Dans la première partie nous introduisons un état de l'art des modèles de veille existants. Par la suite, nous enchaînerons la présentation de notre approche et de sa mise en œuvre. Et enfin nous conclurons le papier par une discussion des résultats et des perspectives.

1 Introduction

La veille consiste à analyser l'information dont on dispose dans le cadre d'un processus continu. Elle permet d'obtenir des informations facilitant la prise de décision tout en réduisant les risques liés à l'incertitude. Les premiers modèles de veille ont été inspirés du cycle de renseignement stratégique en termes militaires proposé par le Sénat des États-Unis en 1976. Ce modèle fonctionne sur le mode des questions et réponses [6]. L'AFNOR[1] a proposé un modèle type de veille dont lequel le processus passe par huit étapes principales [1]. Alors que [7] a repris les phases principales du modèle de l'AFNOR auxquelles il a intégré trois autres phases complémentaires qui lui apparaissent intéressantes et essentielles à l'amélioration continue du processus de veille. L'équipe de recherche Lesca[2] a construit une méthodologie aidant à mettre en place le dispositif de veille stratégique appelée la méthode L.E.SCAnning [12].

[1] Association Française de Normalisation, norme X50-053
[2] http://veille-strategique.eolas-services.com/equipe/equipe.htm

Ces modèles et autres ont pour objectif de facilité l'intégration formelle de la veille au sein des organisations. Pourtant, en pratique il n'existe ni procédé ni outil capable de réaliser toutes les opérations du processus de veille. Hors, suivant le type de sources d'information, qu'il soit hors ligne ou en ligne sur Internet plus précisément, le processus de veille utilise des procédés et des outils distincts. Nous présentons dans ce document une méthode pratique d'analyse de documents textuels pour la veille technologique en télécommunication. Cette méthode a été validée par une équipe de veilleurs dans le cadre d'un projet de Maroc télécom.

2 Méthode proposée

2.1 Vue d'ensemble

Nous nous somme basé sur Internet comme ressource d'information du fait qu'il est considéré comme étant une source en ligne publique et variée. Les organisations lui portent plus d'intérêt dans la pratique de la veille en analysant les sources qu'il détient, en particulier les documents textes. Selon la véracité de l'information publiée et visée sur Internet, plusieurs études de cas [5], [15], [16], [17] et autres, ont montré l'impact de l'Internet sur l'efficacité et la qualité des différentes phases de la veille.

Le processus passe en générale par six phases principales (cf. figure 1) : le ciblage des sources, la collecte, la surveillance automatisée des sources, l'extraction et la catégorisation de concepts, la modélisation et la catégorisation des documents et la phase de synthèse et diffusion.

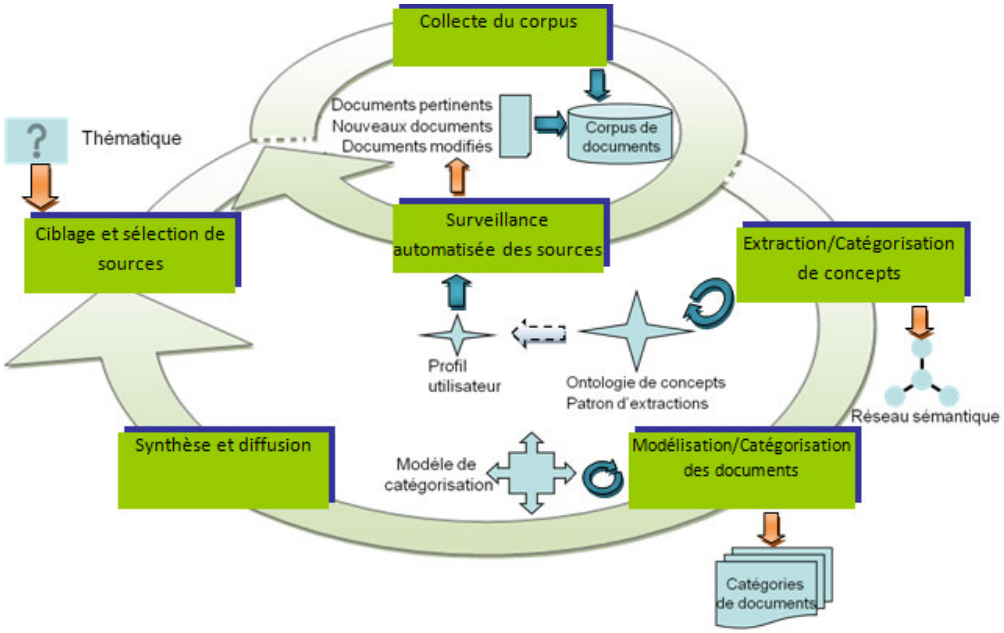

Figure 1 - Approche pratique de la veille technologique sur le web

2.2 Ciblage et sélection de sources

Avec les objectifs de la veille technologique en tête [8], nous avons configuré puis lancé plusieurs requêtes, composées de termes bien choisis, avec les moteurs de recherches connus (Google, Bing, Yahoo,…). Les liens pertinents ont été intelligemment rassemblés. Les sites web des opérateurs ainsi que les sites des entreprises, des administrations publiques et des organisations actives dans la technologie et le secteur de télécommunication ont été consultés. Sans oublier des sites web couramment consultés et considérés favoris par une équipe de veilleurs Maroc Télécom.

2.3 Collecte du corpus

Lors de la première itération de la méthode proposée, la phase de collecte du corpus reste fastidieuse et semi-automatique. Les documents composant le corpus sont soigneusement collectés à la main à partir des sources sélectionnés précédemment. La contrainte de base est que les documents doivent contenir un maximum de concepts liés à la thématique étudié.

L'outil utilisé pour l'extraction de concepts (voir section 2.4) utilise des ressources linguistiques dédiées au traitement automatique des langues (TAL). Ces ressources sont composées généralement de dictionnaires, thesaurus et d'ontologies de concepts. C'est la raison pour laquelle la première itération de la collecte doit aboutir à un corpus riche en vocabulaire du domaine afin de construire et/ou d'adapter ces ressources linguistiques à la thématique étudiée.

2.4 Extraction et catégorisation de concepts

L'extraction est réalisée par l'outil Text Mining for Clementine (TM4C) de SPSS[3]. TM4C procède à une extraction automatique des concepts et leurs fréquences à partir des fichiers textes du corpus. Néanmoins, cette extraction ne répond pas souvent aux attentes des veilleurs; tel que la liste des concepts extraits nécessitent un affinement supplémentaire. Par exemple, Il existe plusieurs concepts extraits et qui indiquent la même chose dans la problématique étudiée. Dans ce cas, un groupement de ces concepts est nécessaire dans un seul concept clé : les concepts « MMS » et « SMS », pour notre cas, peuvent être groupés et catégorisé dans le concept clé « messagerie ».

Dans un travail précédent [13], nous avons présenté le processus d'enrichissement et d'affinement des ressources linguistiques de l'outil utilisé. C'est un processus itératif, semi-automatique qui aboutie à l'élaboration d'une ontologie de concepts du domaine étudié. Des travaux, [2], [11], [14],[3] et [4] ont déjà été établis dans ce sens. Ils se sont basés sur l''hypothèse que les connaissances, et donc les concepts permettant de modéliser un domaine donné, sont contenus dans des corpus de textes représentatifs du domaine.

2.5 Surveillance automatisée des sources

Pour que le processus de veille en ligne itère de façon continu, la collecte et l'analyse des documents, provenant des sources cible, ne doivent s'achever. Hors, comme on vient de le citer avant, la collecte demande des interventions humaines considérables, ces dernières consistent en la recherche, la visualisation et la validation de documents candidats à placer dans le corpus. Pour cela, nous avons conçu un outil automatisant cette étape de recherche, de surveillance et de collecte de corpus.

Après avoir choisi les sites web à surveiller, l'utilisateur configure sa requête composée de concepts appartenant à l'ontologie basique déjà crée (section 2.4). L'outil feuillette en temps réel et indexe les pages HTML des sources cibles. En sortie, il met à jours une base d'indexation des documents pertinents et renvoie à l'utilisateur la liste de documents pertinents comme réponse à sa requête (cf. figure 2). La surveillance se limite à détecter les nouveaux documents pertinents et les documents modifiés parmi ceux anciennement indexés. Tous ces documents peuvent enrichir le corpus de documents pour une éventuelle itération du processus de veille et d'enrichissement de l'ontologie.

[3] www.spss.com

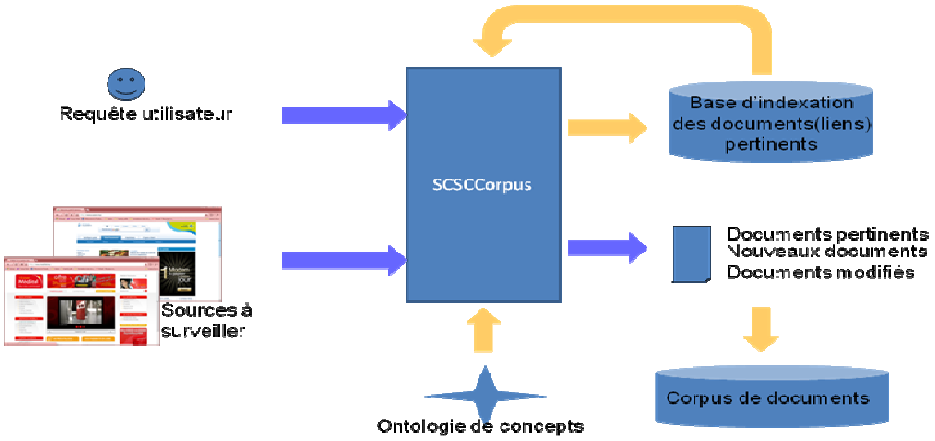

Figure 2 - Schéma fonctionnel de l'outil de surveillance

2.6 Modélisation et catégorisation des documents

La modélisation permet de donner un score pour chaque document puis de décider les classes gagnantes auxquelles le document doit appartenir. Une étude de cas montrant ceci est présente dans un travail précédant [13]. La technique de modélisation utilisée est l'analyse en composante principale (ACP) [9]. Cette dernière ne force pas un document d'appartenir à un et un seul cluster. Les documents ainsi obtenus peuvent être classés selon des facteurs multiples et par conséquent peuvent se trouver à la fois dans plusieurs clusters [10]. D'autre part, les composantes trouvées sont indépendantes et permettant d'améliorer les résultats des techniques de catégorisation (ou classement) sur les facteurs.

2.7 Synthèse et diffusion

Dans la phase de synthèse, il s'agit de dégager le « sens » ou les aspects critiques et stratégiques des informations collectées, notamment les signaux faibles et surtout de proposer une formulation adaptée au processus de décision de l'entreprise. La diffusion quand à elle consiste en la mise à disposition des informations, dans des livrables spécifique et d'effectuer une communication périodique sous des formes diverses : note, bulletin, lettre d'information, dossier… etc. Ces produits ne remplissent leur fonction que s'ils parviennent aux décideurs à temps et sous la forme appropriés. Une phase de validation et de réajustement permet après communication des résultats, un ajustement par approfondissement ou réorientation des objectifs de la veille [1].

3 Discussion

3.1 L'ontologie résultat

Nous sommes en train d'améliorer l'ontologie en analysant les concepts présents dans le corpus. Avec un affinement itératif de la liste de concepts extraits nous augmentons petit à petit la représentation de l'ontologie. L'ontologie contient actuellement 5656 concepts qui se répartissent hiérarchiquement autour de 42 concepts clés. Le niveau élevé de groupement de concepts fournis par les concepts clés permet de voir les tendances générales dans le corpus de documents, même avant de faire toute catégorisation de document. Aussi, La fréquence d'un concept spécifique peut être un indicateur pour savoir si l'ontologie utilisée fait ou pas un bon travail de trouver une grande partie de documents pour ce concept clé.

3.2 Outil pour la collecte et la surveillance automatique de sites web

La plupart des outils et moteurs de recherche de ce genre utilisent des indexes construit et limités au niveau syntaxique des mots. Ainsi, ils retournent les documents pertinents à base d'une simple comparaison de chaines de caractères entre les mots. Hors, ces indexes, vides de relations sémantiques telle que la synonymie, ne peuvent différencier entre les homonymes (« souris » comme périphérique et « souris » comme animal) et des résultats imprévus arrivent à l'utilisateur. D'autre part, des documents peuvent être pertinents mais non renvoyés à l'utilisateur du fait qu'ils ne contiennent que des mots proches sémantiquement des concepts recherchés.

Par une liaison sémantique des mots trouvés dans les documents HTML, à des concepts de l'ontologie, précédemment construite (section 2.4), les documents sont indexés puis affichés par ordre de leur pertinence en réponse au profil de recherche crée par le veilleur. Notre outil utilise l'ontologie de concepts pour supporter la désambiguïté et par conséquent de réduire le bruit dans les documents renvoyés, et aussi pour diminuer le silence en augmentant le nombre de documents pertinent sémantiquement. Un prototype est déjà mis en test, nous prévoyons une amélioration de la recherche et de l'indexation en intégrant le voisinage d'un concept, construit à base des différentes relations élémentaires dans l'ontologie.

4 Conclusion

Dans cet article, nous avons montré une méthode pratique d'analyse de documents texte et de surveillance automatique de site web pour la veille technologique en télécommunication. Sa mise en ouvre consiste en la sélection de sources web à surveiller, un outil a été développé afin de collecter en local les documents pertinents. Puis, le corpus de document est analysé afin de détecter de nouvelles connaissances et de dégager le sens stratégique des informations collectées. Dans la première partie nous avons présenté un état de l'art des modèles de veille existants. Par la suite, nous avons expliqué notre approche et sa mise en œuvre.

5 Bibliographie

[1] AFNOR, (1998), Prestations de veille et prestations de mise en place d'un système de veille, Paris.

[2] AUSSENAC-GILLES ET AL., (2003), *Analyse comparative de corpus : cas de l'ingénierie des connaissances*, Actes de la Conférence Ingénierie des Connaissances, Laval.

[3] BOURIGAULT D., LAME G., (2002), *Analyse distributionnelle et structuration de terminologie, application à la construction d'une ontologie documentaire du droit*, TAL, Vol. 43, nº 1, p. 129–150.

[4] BOURIGAULT D., AUSSENAC-GILLES N., (2003), *Construction d'ontologies à partir de textes*, Actes de la 10ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2003), T2, pp. 27-50.

[5] BRABSTON M-E, MCNAMARA G., (1998), *The Internet as a competitive knowledge tool for top managers*, Industrial Management & Data Systems.

[6] CHAGNEAU V.,(NOVEMBRE 2006), *L'information stratégique dans l'intelligence économique et la stratégie : émergence d'une logique heuristique*, actes du colloque intelligence économique et compétition internationale, Paris.

[7] DEROUET D., LEPOIVRE F., (2005), *veilles, processus et méthodologie*, NEVAOCONSEIL.

[8] GERARD VERNA (NOVEMBRE 1993), *La veille technologique, une ardente nécessité*, Université Laval.

[9] HAIR, J.F. ET AL, (1992), *Multivariate data analysis*, (3rd ed.), New York: Macmillan.

[10] KONGTHON A., (AVRIL 2004), *A Text Mining Framework for Discovering Technological Intelligence to Support Science and Technology Management*.

[11] LAME G., (2000), *Acquisition de connaissances à partir de textes, vers l'élaboration d'une ontologie du droit*, Actes RJCIA 2000, p. 211–221.

[12] LESCA, H. ET KRIAA, S., (janvier 2006), *Veille Anticipative Stratégique : vers une gestion de la connaissance tacite dans les PME-PMI*, Séminaire VSST'2006, Veille Stratégique, Scientifique et Technologique, Lille.

[13] MACHHOUR, H., EL HIMDI, K., BERRADA, I., KASSOU, I., (2007), *Veille et extraction de concepts: étude de cas dans le domaine des télécommunications*, colloque Veille stratégique scientifique et technique (VSST), Marrakech.

[14] NAVIGLI R., (2002), *Automatically Extending, Pruning and Trimming General Purpose Ontologies*, International Conference on Systems, Man and Cybernetics, p. 631– 635.

[15] REVELLI C., (1998), *Intelligence stratégique sur Internet*.

[16] TEO T.S.H., (2000), *Using the Internet for competitive intelligence in Singapore* , Competitive Intelligence Review.

[17] TEO T.S.H., CHOO W. Y., (2001), *Assessing the impact of using Internet for competitive intelligence*, Information Management.