L'analyse de l'information proposée par les outils de veille

Claire François

L'analyse de l'information proposée par les outils de veille

Claire François

Unité Recherche et Innovation(URI), INIST/CNRS,
2, allée du Parc de Brabois, 54514 Vandoeuvre-lès-Nancy, France
claire.francois@inist.fr

Introduction : le processus de veille

La veille est communément définie comme un processus intégrant les phases de définition du besoin de veille, de recherche et collecte des informations brutes, de traitement et analyse afin de transformer ces dernières en information utile, et enfin de diffusion sélective pour l'aide à la prise de décisions au sein de l'entreprise. Cette dernière étape conduit à affiner la définition des besoins dans une dynamique interactive entre les différents acteurs de la veille.

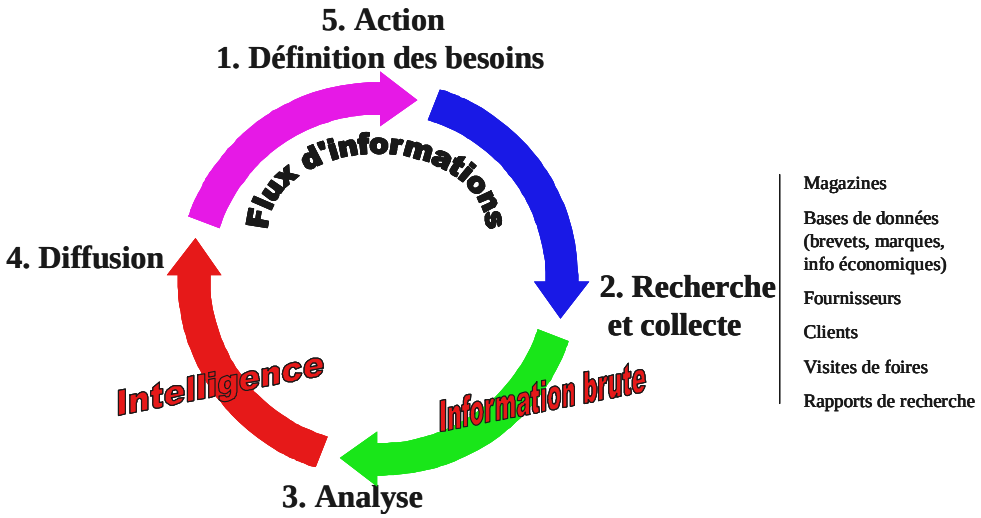

Figure 1 : flux d'informations pendant le processus de veille.

L'aspect dynamique de la recherche d'information est essentiel et intègre donc la notion de surveillance des différentes sources sélectionnées afin d'obtenir l'information la plus à jour possible. Le volume de l'information collectée rend nécessaire la définition d'une méthodologie de gestion de cette information afin de garder accessible toute information utile. La collecte permet de repérer une information spécifique qui, si elle passe inaperçue, peut avoir des conséquences graves pour l'entreprise. Cependant, une information seule peut ne pas être directement utile, le recoupement entre plusieurs informations permet de comprendre des événements dans l'environnement de l'entreprise, et repérer des tendances. C'est le rôle de l'analyse de l'information, elle requièrt une expertise du domaine et peut être soutenue par une méthodologie intégrant les aspects d'analyse textuelle des documents collectés, de catégorisation permettant une navigation plus aisée, de bibliométrie (mesure du flux de l'information), analyse des réseaux, de classification et de cartographie, et enfin de suivi temporel. La visualisation sous forme graphique de l'information permet d'utiliser la puissance de ce mode de communication en complément de la forme textuelle des données à analyser.

Collecte et gestion de l'information

Outre la collecte des informations factuelles, la recherche d'information sera réalisée sur des bases de données de brevets, données économiques, bibliographiques,… L'information disponible sur ces bases présente l'avantage d'être validée et structurée ce qui permet un traitement ultérieur automatisé. La recherche d'information sur Internet se répartit entre la navigation sur des annuaires et portails sélectionnés par le veilleur (ex : Yahoo, Voilà ou Excite), et l'utilisation de moteurs ou agents de recherche qui permettent de définir des requêtes plus ou moins complexes et de les appliquer sur les sites Web (ex : Google, Altavisa, Lycos, Exalead, Copernic ou Intution/iInternet de Sinequa). Pour une information plus complète sur les agents voir le site www.agentland.fr.
Les portails et moteurs proposent souvent des fonctionnalités de surveillance des sources définies par l'utilisateur : sites spécifiques, listes de discussions… Il existe également beaucoup d'agents d'alerte réalisant cette fonctionnalité (ex : AMI Market Intelligence, Website Watcher, KBCrawl, Wysigot, …).
De nombreux outils de gestion documentaire ou de gestion de connaissances peuvent être utilisés pour gérer la masse d'information recueillie, ces outils permettent le stockage, l'indexation, la catégorisation des documents et la capitalisation des connaissances. Enfin, les systèmes complets de gestion de l'information (de type Digimind Evolution, Keywatch, Aperto Libro, KnowledgeManager, KB Crawl, Kaliwatch Professional, Pericles, Verity K2 entreprise, Lingway KM, Tropes Zoom, Autonomy VS 2000, …) intègrent ces différentes fonctionnalités de collecte, gestion et diffusion de l'information.

Analyse textuelle de l'information

L'analyse textuelle peut être effectuée sur les requêtes afin d'améliorer la phase de collecte ou les textes collectés dans la phase de traitement. Dans ce dernier cas, elle permet une indexation des textes qui sera ensuite utilisée pour catégoriser, classer ou cartographier des documents collectés. Cette analyse textuelle est réalisée selon une approche statistique ou linguistique, ou en combinant les deux.

Les approches statistiques intègrent les traitements de lemmatisation des mots du texte (ou termes) en les ramenant à leurs racines, les calculs de fréquences et cooccurrences des termes dans les textes permettant de repérer les associations importantes entre concepts, le calcul des **s**egments répétés (suites de termes que l'on retrouve fréquemment dans les textes qui identifient des concepts plus précis et riches qu'un mot simple). Afin d'améliorer la précision des résultats, ces approches incluent l'utilisation de ressources terminologiques de type «dictionnaires de mots vides», ou «dictionnaires d'équivalence», ainsi que des modules linguistiques. Les outils basés sur ce type d'approche sont par exemple Alceste, Tétralogie, WordMapper, Sphynix Lexica, Périclès.

Les approches linguistiques regroupent l'analyse morpho-lexicale qui reconnaît les « lemme » et des flexions associées des mots, l'analyse syntaxique qui établit un schéma des relations entre les mots en se basant sur une grammaire de la phrase et enfin l'analyse sémantique qui construit une représentation conceptuelle du sens du texte. Cette dernière analyse nécessite des ressources terminologiques élaborées modélisant des domaines très ciblés. L'indexation des textes peut être réalisée soit par extraction des descripteurs potentiels, soit par reconnaissance de termes présents dans des ressources terminologiques. Dans ce cas, les termes retrouvés dans les textes sont présents soit sous la même forme que dans la ressource soit sous des formes variantes. Ces variations peuvent être flexionnelles

(identification des formes singulier / pluriel des noms), morpho-derivationnelles (dérivation de mots en d'autres mots de catégories grammaticales différentes, par exemple dérivation d'un adjectif à partir d'un mot), et enfin syntaxiques (insertion d'un mot ou d'une forme coordonnée à l'intérieur d'un groupe nominal, permutation de groupes de mots autour d'un élément pivot). La recherche des entités nommées (noms de personnes, de lieux, d'entreprises, dates, unités monétaires, pourcentages) fait intervenir des grammaires spécifiques associées à des marqueurs linguistiques, et également des référentiels. Certains outils proposent également des fonctionnalités d'interrogation multilingue, de résumé ou traduction automatique.
Les outils utilisant ce type d'approche sont par exemple Lingway KM, Kaliwatch Professional, Pertimm, Tropes, Insight Discoverer Extractor, Lexiquest Mine, Stanlayst. Pour compléter voir le site de l'APIL (Association des Professionnels des Industries de la langue).

Catégorisation des documents collectés

La gestion des documents collectés passe par une catégorisation qui sera réalisée par affectation des documents dans une arborescence selon un mode supervisé, c'est-à-dire que l'arborescence est prédéfinie (ex : Exalead, Insight Discoverer Categorizer, LexiQuest Categorize, Lingway KM, Pertimm, Kaliwatch Professional,…). La méthode la plus employée est un apprentissage statistique qui peut être schématisé de la façon suivante : une fois le plan de classement validé, des documents déjà catégorisés sont utilisés comme base d'apprentissage pour le moteur de catégorisation, qui pourra ensuite affecter automatiquement les nouveaux documents dans le plan de classement par mesure de similarités avec les documents déjà catégorisés. La catégorisation peut également être réalisée par une analyse linguistique en se basant sur un référentiel terminologique. Ces catégories seront affichées sous forme de listes permettant une navigation dans les documents récoltés.

Analyse bibliométrique de l'information

L'analyse bibliométrique permet d'obtenir une vue globale de l'information collectée par mesure des différentes informations factuelles associées au texte. Les analyses possibles sont :

- le réseau des acteurs par le traitement des auteurs et de leurs affiliations,
- les moyens de communication par analyse des types de publications (revues, rapports, congrès), de leur répartition selon les pays éditeurs pour les publications scientifiques ou selon les pays de protections pour les brevets,
- l'évolution de l'activité de différents domaines par analyse des dates de publication ou de dépôts (pour les brevets).

Les outils permettant ces analyses (ex : Intellixir, Tétralogie, WordMapper, MatheoAnalyser, Serv'IST, Stanalyst) offrent aux utilisateurs la possibilité de construire des tableaux simples et croisés de répartition des références selon un ou plusieurs critères, ainsi que différents modes de représentation graphique des résultats.

Similarités et visualisation sous forme de réseaux

L'analyse du contenu des documents collectés a pour but de détecter une organisation des connaissances enfouies dans les textes permettant de faire émerger l'information utile pour les décideurs.
Les différentes méthodes développées se basent sur la définition d'une similarité entre documents, pour cela il est nécessaire dans un premier temps de représenter chaque document par un vecteur de termes, ces derniers sont le fruit d'une indexation qui peut préexister, être

ajoutée par l'utilisateur au moment de la collecte, ou enfin être ajoutée par l'analyse textuelle présentée ci-dessus. Pour les sites Web, cette similarité peut être calculée à partir des liens hypertexte.
L'ensemble de ces similarités définit un graphe ou réseau qui sera directement affiché à l'écran (ex : Lexiquest Mine, Wordmapper, Tétralogie, BibTechMon), ou pour les moteurs : Webbrain, kartoo, Mapstan, Mooter). Ces graphes peuvent représenter directement les relations entre pages collectés pour les moteurs de recherche, les relations entre documents collectés ou entre informations extraites des documents et analysées en temps que tel (termes d'indexation, auteurs, …), ils permettent une analyse et améliorent la navigation dans les données récoltées.

Classifications automatiques et cartographies

Les classifications automatiques (cluster analysis ou clustering en anglais) groupent directement les documents en classes, en se basant cette fois-ci directement sur leur similarités. Parmi les différentes méthodes existantes, nous pouvons distinguer des méthodes de classification hiérarchique dont le résultat est une hiérarchie de classes qui sera directement affichée à l'écran (ex : Vivisimo). Par opposition, les méthodes de classification non hiérarchique permettant de calculer directement une partition. Les résultats seront présentés sous forme d'une liste de classes (Lingway KM), d'un réseau de classes (ex : Insight Discoverer Clusterer), d'un graphique selon indice de centralité et de densité (ex :WordMapper, Stanalyst). D'autre part, les analyses factorielles produisent des représentations graphiques permettant de positionner les objets analysés sur un plan et définissent une carte qui sera affichée (ex : Tétralogie, Stat'Mania associé à Tropes). Ces méthodes sont utilisées en complément des classifications afin de cartographier les classes obtenues (ex : Tétralogie, CartoWeb, Stanalyst). Une méthode neuronale comme les cartes auto-adaptatives de Kohonen produit des classes directement positionnées sur une grille (ex : Websom), cette grille sera donc affichée à l'utilisateur.
Différents outils comme Tétralogie, Clémentine, MatheoAnalyser, Stat'Mania, Périclès, Miner 3D, Analyst's Netebook ou Stanalyst proposent l'application de différentes méthodes de classification et de visualisation sur un même corpus de textes afin de réaliser une analyse approfondie de l'information collectée.

Analyse temporelle de l'information

L'analyse temporelle des informations collectées peut être réalisée sous la forme d'une courbe de suivi dans le temps d'un sujet ou d'un terme (ex : Péricles, Intellixir, Lexiquest Mine , AMI Market Intelligence, Analyst's Netebook). Quand une classification des documents est réalisée, un outil comme WordMapper permet de réaliser 2 classifications à un temps T et T+1 et de calculer les différences observables entre ces deux résultats, tandis qu'un outil comme Calliope propose de distinguer parmi des classes obtenues des thèmes émergents, stabilisés et déclinants.

Conclusion

L'ambition de cet exposé n'est pas de donner une liste exhaustive de tous les outils existants dans un marché très évolutif et innovant, mais plutôt de dresser un panorama des différentes approches mises en place pour analyser l'information collectée, illustré par quelques exemples. Pour des informations complémentaires, voir aussi le benchmarking en cours à l'adresse : http://veille-srv.inist.fr/~OutilsVeille/Public/.