Le marché mondial du Text Mining

panorama de l'offre, évaluation quantitative, tendances et perspectives d'évolutions.

Par Philippe Bonny et Alain Garnier

Selon le cabinet IDC, la production mondiale d'information (en volume) dans les bases structurées n'a augmenté que de 4% en 2006 et ne représente qu'environ 10% des données, a contrario, les données non structurées représentent quant à elles 90% du total et augmentent de 6400% par an ! Le traitement des données non structurées constitue donc un enjeu colossal pour aujourd'hui et encore plus pour demain. Les logiciels et technologies dits de « Text Mining », encore en émergence, constituent les premiers leviers d'exploitation de cette manne informationnelle. Comme peu ou pas d'études de marché sur ces outils ne sont disponibles aujourd'hui, Inevidence a cherché à combler cette absence en dressant de manière synthétique un panorama global de ce marché.

Le moteur de recherche est la deuxième fonction la plus utilisée après le mail dans l'entreprise! Et pour cause, le « search » constitue le process automatique qui permet de fournir un accès rapide et fluide à une masse d'informations toujours plus grande et toujours plus non-structurée pour ensuite le restituer de manière utile et scénarisé à l'utilisateur.

Au delà du Search, les enjeux

Le premier étage de la technologie a été de donner simplement la liste des documents qui répondent à un mot clé. Cette première génération de moteur « full text » a marqué son temps. Puis, les moteurs évoluant, ils ont été capables de présenter les « méta-données » associées au document : le titre, la date, l'auteur, la source, gérer la sécurité… Ensuite, Google a imposé définitivement l'utilisation de l'extrait pertinent (ou « snippets ») qui permet de voir en deux lignes si le document retourné correspond à la recherche.

L'étape naturelle suivante est donc que ce processus d'enrichissement des informations produites par le moteur de recherche s'appuie désormais sur les technologies de Text Mining. Tout d'abord localement, sur le document, afin d'extraire des données plus fines comme les personnes, les acteurs, les lieux, les actions (rachat, fusion, annonce, vente…) ; Mais aussi globale au niveau du corpus afin de faire ressortir des thèmes récurrents, des problématiques émergentes etc…

Le Text Mining apporte donc au delà du search un début d'analyse du corpus qui vient considérablement enrichir l'expérience utilisateur. Le couple technologique/société formé par les moteurs de recherche d'une part et les acteurs du Text Mining d'autre part n'a pas fini son rapprochement de plus en plus serré.

La valeur ajoutée du Text Mining

Dans les années 60 sont nées un ensemble de technologies qui visaient à exploiter les données sous formes tabulées (volume de vente, âge, fréquence d'achat, lieu, etc.), technologies que l'on a appelées « Data Mining ». Il s'agissait de reconnaître des familles de données proches (ou au contraire éloignées), d'identifier des relations entre celles-ci (cause, conséquence, association forte, ...), de repérer et de caractériser des tendances (croissance, décroissance, forme d'évolution ...), de construire des modèles de prédiction (comportement, ...). Aujourd'hui ces technologies sont matures, largement diffusées et sont appliquées sur de grandes bases de données tabulées (données économiques, bases de données clients, données de production, etc.).

A la toute fin du siècle dernier, avec l'avènement d'Internet notamment, sont nées de nouveaux ensembles de technologies qui visaient à étendre le champs d'exploitation des données aux données non tabulées et éparpillées dans les textes, technologies que l'on appelle aujourd'hui « Text Mining ».

Valeur ajoutée du Text Mining

Il s'agissait là aussi de reconnaître dans des corpus, des éléments textuels proches (acteurs, thèmes, ...), d'identifier des relations entre ces éléments (entre acteurs, entre thèmes, entre acteurs et thèmes, etc.), de repérer des tendances (teneur du discours, sentiments positifs / négatifs, etc ...), et enfin de construire là aussi des modèles prédictifs sur la base d'éléments identifiés dans les textes (ex. en chimie/biologie « A inhibits B » and « B activates C » then « A (might) inhibits C »). Les technologies de Text Mining sont aujourd'hui plus matures et sont appliquées sur des corpus textuels très divers, qu'il s'agisse de pages web, de blogs, de questionnaires ouverts, de transcripts de call center, d'Intranet d'entreprise, ou encore de base de presse par exemple.
Ces technologies utilisent, au delà de modèles statistiques et mathématiques communs avec le Data Mining, des connaissances spécifiques liées à la linguistique (analyse lexicale, morphosyntaxique, etc.) et à la sémantique entre autre. En outre, elles doivent s'affranchir des langues et développer des processus multilingues ou cross lingues.

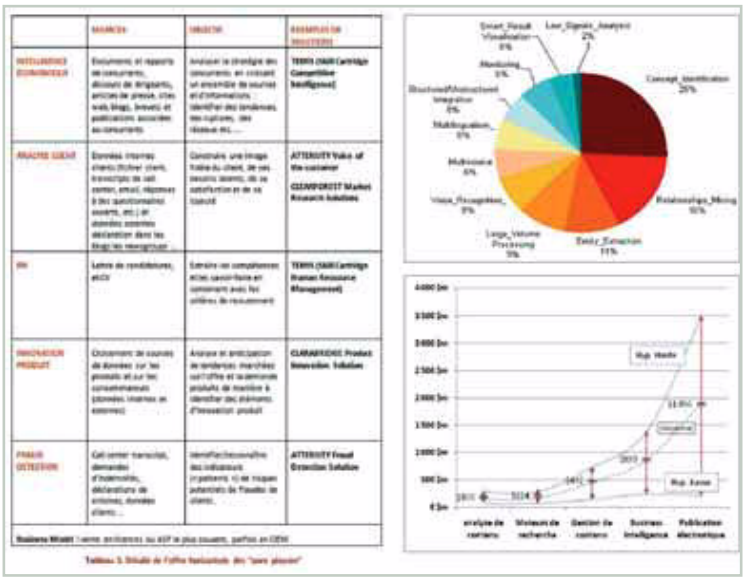
En somme, on peut caractériser simplement la valeur ajoutée des logiciels et technologies de Text Mining comme un processus de traitement de corpus de documents, créateur de gain de temps et d'intelligence pour l'exploitation de ces documents

L'étude de marché Inevidence

L'objectif de cette étude est de dresser un panorama global du marché du Text Mining au niveau mondial et de dresser les perspectives principales quant à son évolution future. Dans un premier temps, une analyse de l'offre dans le monde par grands segments de marché est effectuée. Dans un deuxième temps, une estimation du chiffre d'affaire généré segment par segment est présentée ainsi qu'une étude prévisionnelle de leur évolution à l'horizon 2011. Ensuite une approche plus qualitative des tendances du marché est décrite : applications clés, sources utilisées et traitements employés. Enfin les perspectives principales d'évolution du Text Mining sont exposées. L'étude se clôture par un focus sur les acteurs français, leur positionnement et leurs résultats face à ce marché...

Inevidence est un cabinet d'études et de conseil expert en traitement avancé de l'information crée en 2005 par Philippe Bonny et Alain Garnier.

Il propose des prestations d'études, de conseil et de formation notamment autour des problématiques d'intelligence marketing, d'intelligence économique, d'intelligence média ou d'innovation. Voir le site d'Inevidence www.inevidence.fr.



1 Selon cette étude, le marché mondial du Text Mining s'est élevé globalement à 1,2 milliard de dollars en 2006 en croissance de 21% par rapport à l'année 2005 et devrait réaliser une croissance annuelle moyenne (CAGR) de 24% jusqu'en 2011 où il devrait atteindre 3,6 milliards de dollars dans le monde.

2 Une dizaine d'acteurs « pure players » dans le monde : Attensity (US), Basis Technology (UK), Clarabridge (US), Megaputer (US), NStein (Ca), Temis (Fr), Teragram (US). Deux acteurs leaders ont été acquis en 2007, ClearForest (Israël) par Business Object et Inxight(US) par Reuter. Une pléthore de petits acteurs nationaux dans le monde, en France par exemple : Image, Acetic, Semiosys etc. Des acteurs non « pure players » de plus en plus nombreux à intégrer cette technologie dans leur offre (éditeurs de moteurs de recherche, gestion de contenu, business intelligence, publication électronique, ...).