Text Mining pour l'intelligence économique et l'analyse stratégique : la solution mise en place dans un grand groupe industriel.

Anne BONNET-LIGEON (TOTAL SA , 2 place de la Coupole, 92078 Paris La Défense cedex, France)
anne-genevieve.bonnet-ligeon@total.com

Mots clefs :
Fouille de données textuelles, catégorisation, analyse stratégique, veille concurrentielle, veille stratégique, intelligence économique, analyse de données

Keywords:
Text-mining, categorization, strategic analysis, business intelligence, environmental scanning, competitive intelligence, data analysis

Palabras clave :
Explotacion del texto, clasificacion, analisis estratégica, vigilancia estratégica, inteligencia competitiva, analisis dato.

Résumé

La Division en charge de l'information au sein de la Holding a pour mission de fournir de l'information ciblée de qualité aux différentes entités du groupe et de mettre en place des processus opérationnels de collecte et d'analyse.
L'accroissement des volumes d'information à traiter, dû à la quantité des sources disponibles (géopolitique, technique, environnement, finances, social, micro et macro-économique, …) et la diversité des demandes internes, rend les traitements manuels de lecture et la synthèse difficiles. Il était urgent d'automatiser et de rendre rapide l'accès à l'information.

La solution de text-mining mise en place optimise les informations issues des sources de presse, des périodiques disponibles sur le portail du groupe, des communiqués de presse, des documents issus du web, ou des documents produits en interne. C'est pour traiter en continu cet ensemble d'informations que la solution Insight Discoverer™ Extractor associée à la Skill Cartridge™ Competitive Intelligence de TEMIS a été mise en place. Elle permet d'extraire instantanément et en continu de ces flux des données concernant les investissements, les prises de participation, les indicateurs financiers, les parts de marché, les partenariats, les infrastructures, les organigrammes concurrents, les acteurs, …
Cette information est mise à disposition quotidiennement auprès des utilisateurs via l'intranet du groupe.

1 Introduction

Face au leitmotiv « les utilisateurs n'ont pas le temps de lire » , il faut rendre l'accès à l'information plus attractif, attirer l'attention sur ce qui intéresse, faire ressortir les mots importants, les catégories d'information.

L'objectif du projet est donc de fournir aux utilisateurs la possibilité d'exploiter réellement la masse d'information mise à disposition quotidiennement sous forme électronique, afin de faciliter l'analyse stratégique et la prise de décision.

Pour cela, il est nécessaire que l'information utile extraite leur soit mise à disposition par le biais d'une application dont l'usage est généralisé et requérant un nombre de manipulations faible : l'intranet du groupe.

2 Cadre du projet

La Division en charge de l'information pour l'ensemble du groupe à la holding a pour mission de faire en sorte que toute structure du groupe Total dispose des informations qui lui sont nécessaires pour comprendre son action, se situer dans son environnement économique et concurrentiel, faire valoir ses droits, préparer son avenir, travailler plus vite et mieux.

Les thèmes traités dans le cadre de veille stratégique et concurrentielle sont très semblables quel que soit le domaine ou l'entité du groupe. Il s'agit de surveiller les sociétés concurrentes dans leurs activités, les parts de marchés entre acteurs, fusions-acquisitions, politique de prix, ...

La Division a par ailleurs pour objectif la mise en oeuvre d'une application de fouille de données textuelles, « text-mining », dans le cadre de l'exploitation et la valorisation des produits d'information qu'elle diffuse au sein du Groupe.

L'extraction d'information se situe en aval de tout système de diffusion d'information sous forme électronique (revues de presse automatisées, portail de diffusion d'information en texte intégral, etc…)

La réduction conséquente du temps passé à la lecture et l'exploitation des documents pour chacun permet d'envisager le calcul d'un retour sur investissement sensible.

3 Analyse de l'existant

3.1 Description des processus métier et des traitements informatiques

Les services d'information et de documentation de la Division diffusent auprès de leurs clients un grand nombre d'information textuelles, sous forme électronique : toutes les sources disponibles à travers le portail de périodiques en ligne, les documents de réponses aux demandes ponctuelles d'information, des revues de presse, …

D'autres entités du groupe reçoivent, hormis les produits de la Division, des lettres d'informations, revues de presse issues d'organismes extérieurs, rapport et synthèses de banques, etc, largement diffusés.

Ces documents sont lus, dans la plupart des cas, pour en extraire une part d'informations relatives aux activités et projets en cours. Très fréquemment, les informations sont « coupées/collées » ou saisies manuellement par leurs lecteurs dans des fichiers personnels ou partagés, bases de données, selon des classements semblables (sociétés, prix, infrastructures, parts de marché, …).

Par ailleurs, différents produits d'informations sont actuellement réalisés de manière « manuelle » par le personnel de la Division; ceci les limite en nombre et ne permet pas de satisfaire l'ensemble de la clientèle de la Division.

Par exemple, pour le cas des revues de presse, une fois le sujet défini avec le client, la procédure est la suivante :

- connexion à un serveur de presse ou de bases de données externe,
- saisie d'une requête ou consultation de « pistes », procédure de suivi automatique d'un sujet chez le fournisseur,
- sélection et téléchargement des documents jugés pertinents par le documentaliste,
- mise en forme des données téléchargées,
- transmission au client sous forme d'un fichier .doc, rtf, .pdf.,
- lecture « linéaire » plus ou moins approfondie de l'ensemble du fichier par le client,
- repérage d'informations intéressantes par le lecteur, copier-coller ou re-saisie manuelle dans une base de données, tableur, ...

Remarque : cette même information va souvent alimenter x sources personnelles créées par x utilisateurs.

3.2 Défauts du système existant

L'expérience acquise au sein des différents services de documentation et de veille, et les retours sur produits d'information transmis par les utilisateurs finaux (ou clients) ont mis en évidence les faits suivants :

- le nombre de documents et produits d'information mis à disposition des utilisateurs via l'intranet, le web, ou messagerie électronique est en constante augmentation. Il leur est de plus en plus difficile de tout lire et surtout d'accéder aux informations pertinentes indispensables dans le cadre de leurs études.
- la sélection d'information par les veilleurs et documentalistes, dans le but de réduire le volume diffusé entraîne bruit et silence ; la recherche elle-même par des opérateurs booléens n'est pas d'une efficacité très fine sur du plein texte; de plus, la sélection « manuelle » au sein des résultats de ces recherches contient une forte part de subjectivité liée à la non-expertise du veilleur par rapport au sujet traité.
- ce mode de fonctionnement par diffusion sélective de l'information très ciblée et dans un souci de réelle valeur ajoutée contraint le veilleur à limiter le nombre de sujets traités au détriment du service rendu à l'ensemble de ses clients.

- D'autre part, cette situation amène les clients à chercher et à sélectionner l'information eux-mêmes au détriment de leur travail d'analyse, ou pire à se passer d'informations essentielles.
- La lecture « linéaire » rapide d'importants volumes de documents laisse un certain nombre d'informations passées inaperçues.
- Un nombre important de personnes lisent les mêmes documents, pour en retenir les mêmes informations et les sauvegardent souvent, de différentes manières, pour des usages proches (ex. analyse de la concurrence).

4 Mise en place de l'application

4.1 Objectifs

Dans le contexte de surveillance de la concurrence et de l'intelligence économique, il ne s'agit plus seulement de donner du texte brut à nos clients par le biais de tous les systèmes informatiques existant (portail, répertoires partagés, e-mail, etc....) mais d'apporter des réponses précises et directes aux questions qu'ils se posent, par l'extraction automatique de l'information, la visualisation sous une forme simple et conviviale (page web) et la possibilité de la conserver.

Dans le cadre de la lecture et l'analyse de volumes importants de documents textuels :

- mise en place d'un système d'extraction automatique de la connaissance, processus qui permet d'avoir un accès direct à l'information critique contenue dans des textes, en particulier dans le cas de textes non structurés,
- recherche d'automatisation des tâches à la base afin de pouvoir accroître le nombre de sujets traités,
- rendre accessible les résultats via l'intranet selon une interface définie selon plusieurs modes de consultation ; celle-ci peut être déclinée selon les spécificités des unités concernées,
- extraire des informations dans le cadre de la recherche de solutions d'analyse des documents en amont ou en aval des outils déjà utilisés au sein de la Division.

4.2 Pilote de l'application

4.2.1 Méthodologie

Le pilote de l'application a été réalisé dans le cadre du traitement de données habituel pour deux directions du groupe ; des personnes de ces directions ont participé à l'évaluation des résultats de l'extraction automatique.

Deux approches complémentaires de l'extraction d'information des documents on été mises en œuvre :

- d'un part, l'extraction manuelle des informations afin de les valider et de pouvoir comparer les résultats,
- d'autre part, extraction automatisée des informations avec la solution Insight Discoverer Extractor et la skill cartridge CI ,

- en parallèle, définition des interfaces de visualisation des résultats adaptées aux spécificités métiers des utilisateurs .

Le taux de recouvrement des deux approches est de 92%.

4.2.2 Type de documents analysés

Toutes les formes d'informations textuelles rencontrées dans l'entreprise, dans un contexte « intelligence économique » ou veille concurrentielle, sous tout type de format courant : .doc, .rtf, .pdf, .ppt, .html, .xml, …, essentiellement en anglais.

Ce sont essentiellement des dépêches de presse et de fils d'information (Factiva, Reuters, Lexis-Nexis, …), des sources électroniques à texte longs (quelques pages) reçus régulièrement par les directions concernées par le pilote, des documents textes issus du web/internet, des documents issus de bases de données en texte intégral (documents structurés mais avec champ texte long).

L'unité de traitement est le document. L'idéal est de se limiter à des documents dont la taille est inférieure à 2 Mo .

4.2.3 Informations extraites

Une grande partie des informations concernent les concurrents, les fournisseurs de technologie, les fusions/acquisitions, les contrats, parts de marchés, capitalisation, CA, volumes échangés, consortia, joint venture, zones géographiques d'activités, infrastructures, ... ce sont des renseignements identiques collectés par la plupart des entités du groupe.

4.2.4 Caractéristiques des produits participants au pilote

Insight Discoverer Extractor

C'est un moteur d'extraction qui repose sur plusieurs technologies linguistiques développées par la société TEMIS (www.temis-group.com) pour l'analyse grammaticale de documents textuels. Ce serveur comprend tous les documents rédigés dans une des 7 langues suivantes : français, anglais, allemand, espagnol, italien, hollandais et portugais.

L'unité de traitement est le document.

L'extraction d'une information dans un texte comporte quatre phases :

- lecture du texte puis identification de tous les verbes, noms, adjectifs qui le composent.
- détection de la langue rencontrée est réalisée automatiquement par le logiciel ,
- analyse morpho-syntaxique : conversion de tous les mots d'une phrase en leur forme canonique forme non conjuguée,
- pondération, c'est à dire importance relative à accorder aux verbes et aux substantifs

en fonction de leurs occurrences, création d'un vecteur sémantique : normalisation du sens du texte par une méthode attribuant une valeur numérique à chacun des mots.

« Skill cartridge » Competitive Intelligence

ID Extractor fonctionne avec une ou plusieurs cartouche linguistique. Elle fournit les règles à utiliser pour extraire l'information recherchée dans un texte.

La cartouche « Competitive Intelligence » est destinée à la veille économique ou à l'analyse stratégique. Les sources peuvent être des dépêches de presse, site web de news, sites web des concurrents, brevets, publications scientifiques, …

Elle traite essentiellement les données financières, commerciales (parts de marché, nombres de clients, contrats, volumes échangés), boursières (capitalisation, tendances), mais aussi les stratégies d'entreprises (prise de participation, fusion, acquisition, création de joint venture, consortia, nouveaux investissements, …).

Elle est destinée à la recherche d'informations spécifiques, à la constitution des bases de données, et facilite considérablement les analyses stratégiques.

Durant la période du pilote, la cartouche « Competitive Intelligence » a été enrichie par quelques nouveaux concepts (ou règles d'extraction) concernant des aspects propres au domaine pétrolier (infrastructures, unités de mesures pétrole et gaz, investissements …)afin de la rendre particulièrement adaptée aux besoins du groupe.

4.3 Fonctions principales de l'application

L'application (données + logiciels) se trouve sur un serveur dédié. L'accès utilisateurs est sur l'intranet du groupe.

Elle est qualifiée de « consommable » : chaque jour, l'actualité dans les domaines d'information intéressant le groupe est analysée, ainsi que celle des 8 jours précédents. A terme, les extractions seront conservées sur des périodes plus longues.

La plupart des documents étant en anglais , c'est la cartouche « Competitive Intelligence » en langue anglaise qui sera utilisée associée à l'extracteur.

Le déroulement des opérations est le suivant :

- procédures de récupération automatique quotidienne des fichiers de données à traiter sur un serveur
    - ✓ par connexion chez le producteur de données externe (factiva, lexis-nexis)
    - ✓ en interne (contenu des portails),
- possibilité de chargement manuel des données en interne sur le même serveur où à une adresse où le système les prend en compte dans l'analyse,
- le paramétrage de l'extracteur nécessite l'adresse des fichiers sources et un répertoire cible pour la génération des fichiers résultats,
- l'analyseur (extracteur) les récupère, les transforme sous un même format (.xml), les traite et génère des fichiers résultats et des index en .html nécessaire à la visualisation,
- les résultats des extractions sont visualisables sous la forme d'une page intranet ,
- une partie des données extraites est importée dans des tableurs pour leur conservation et/ou un traitement ultérieur,
- des statistiques d'utilisation des données, notamment au niveau de la visualisation d'un article à partir du pointage sur un concept choisi (gestion du droit de copie et de redistribution directement auprès du fournisseur d'informations) sont possibles,

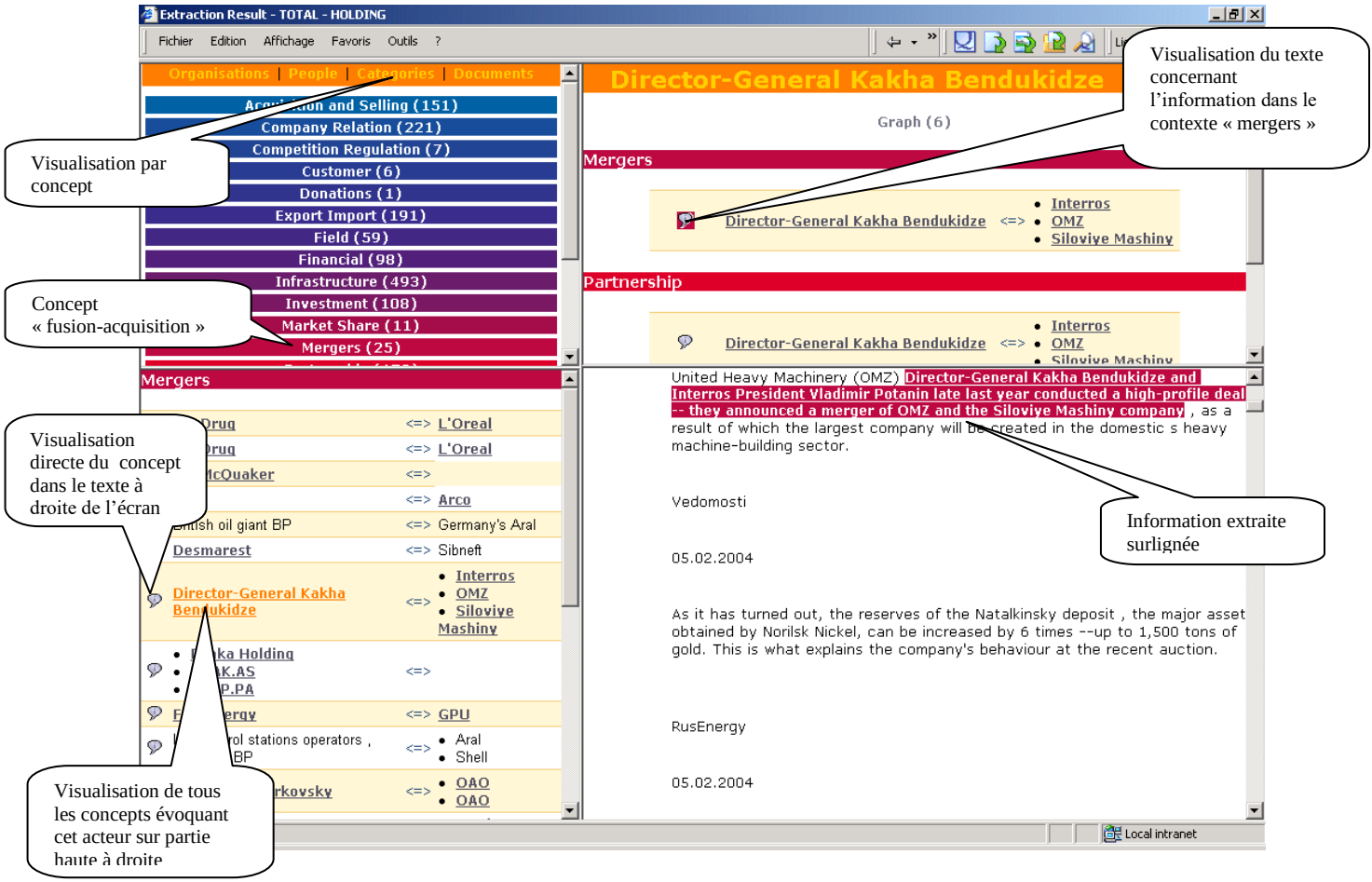
- la mise à disposition de l'information pour une période définie nécessite chaque jour la reprise dans l'analyse de l'ensemble des données cumulées sur cette période.

4.4 Description du contenu de l'interface de visualisation

La visualisation des informations extraites doit pouvoir se faire sous trois formes, chacune étant proposée sur une même première page d'accès comportant la charte graphique Total.

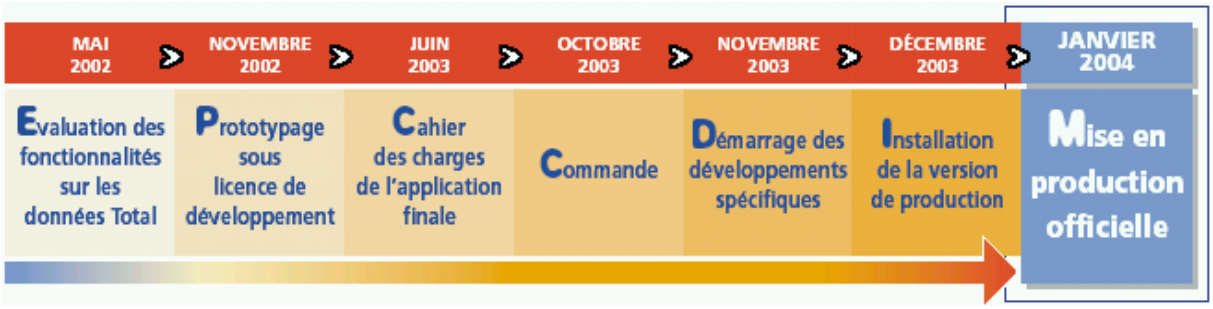
- ✓ Par concept, et au sein de chaque concept par ordre alphabétique d'acteur,
- ✓ Par nom de société ou organisme, et par concept au sein de chacune. Un index alphabétique permet de naviguer parmi les noms de sociétés de manière plus rapide,
- ✓ Par nom de personne, avec un index alphabétique. Ces données sont transférées dans un fichier excel (nom de personne, nom de société, titre, date de l'extraction).

L'interface de visualisation est en anglais.



5 Bilan

5.1 La mise en place : les différentes étapes



5.2 Disponibilité

7 jours sur 7, 24 heures sur 24. Le système est accessible par tout salarié utilisateur quel que soit l'endroit où il se trouve, c'est à dire n'importe où dans le monde.

L'analyse est lancée quotidiennement, à heure fixe. La collecte est régulière, plusieurs fois par 24 heures selon les sources.

5.3 Caractère innovant

Cette application permet un accès direct à l'information en la mettant en valeur directement au sein du document.
Peu à peu les utilisateurs abandonnent la lecture linéaire pour aller directement à l'information utile. Il s'agit réellement de lecture rapide et exhaustive de gros volumes de texte, bien au-delà de nos propres capacités, en se focalisant immédiatement sur les thèmes liés aux activités et préoccupations du lecteur.
La possibilité de naviguer très rapidement au sein de l'information permet de suivre de manière efficace un plus grand nombre de sujets.

Par la capacité d'analyse rapide et performante , cette application constitue un véritable système performant d'aide à la prise de décision.

5.4 Résultats et valorisation économique

Dans l'état actuel de l'application, il y a une diminution sensible (gain supérieur à 50%) du temps de lecture quotidien de chaque utilisateur.
Par ailleurs, l'information est désormais accessible et consultée par des utilisateurs qui n'en avaient pas le temps auparavant

Le système est fiable : la Skill Cartridge™ reconnaît les différents concepts avec précision.

L'optimisation de l'exploitation des informations amène à une optimisation des achats et des coûts pour le groupe.