Veille concurrentielle et veille stratégique : deux applications d'extraction d'information

Bertrand DELECROIX(*)**, Sylvie GUILLEMIN-LANNE**(**)**, Amandine SIX**(**)
bertrand.delecroix@wanadoo.fr , sylvie.guillemin-lanne@temis-group.com, amandine.six@temis-group.com

(*) ISIS/CESD, Université de Marne la Vallée, Cité Descartes – 5 Boulevard Descartes, Champs sur Marne, 77 454 Marne la Vallée
(**)TEMIS, Tour Gamma B, 193-197 rue de Bercy, 75582 Paris cedex 12

Mots clefs :

fouille de données textuelles, ingénierie des connaissances, extraction d'information, règle d'extraction, patron d'extraction, intelligence économique

Keywords :

text mining, knowledge engineering, knowledge extraction, information extraction, extraction rules, extraction pattern., competitive intelligence

Palabras clave :

text mining, ingeniería del conocimiento, extraccion del conocimiento, extraccion de la informacion, reglas de extraccion

Résumé :

Cet article présente l'une des technologies de text mining développées au sein de la société TEMIS, à savoir l'extraction d'information. Il met l'accent sur l'approche adoptée pour le formalisme utilisé et la méthodologie de construction de règles d'extraction d'information par niveaux au sein de la Skill Cartridge™ dédiée Intelligence Economique, couplée au serveur d'extraction d'information. Il est illustré d'exemples concrets issus de deux applications réelles d'intelligence économique :
Le premier exemple concerne une application mise en œuvre pour le groupe France Télécom dans le cadre du projet Extractor. Au sein du service de veille concurrentielle, marketing et financière du groupe, France Télécom a implanté un processus de création et de mise à jour semi-automatique de monographies d'entreprises à partir d'un flux de dépêches Reuters.
Le second exemple concerne la mise en œuvre d'un système d'extraction d'information en intelligence économique déployée au sein d'un organisme de développement économique national. L'ensemble des sources d'information de cet organisme est à ce jour analysé automatiquement, les analystes consultent chaque matin le rapport HTML présentant les informations extraites par le système.

Abstract :

This article focuses on the advantages of organizing rules for knowledge extraction in a hierarchical order. It details the methodology and environment to develop extraction rules, and illustrates the results with examples taken from real-life competitive intelligence applications.
The first example deals with an application implemented for France Telecom Group within the Extractor Project. Within the ARIA, a Business Unit in charge of providing Competitive Intelligence information to the group, TEMIS implemented a process which automatically creates and updates companies' monographs, from a Reuters Business News flow.
The second example deals with the implementation of a business intelligence information extraction system, which was deployed within a public economic development Agency. Each of the Agency information sources are to date automatically analysed. Each morning, an html report is made available to the analysts. It presents the information extracted by the system.

1 Introduction

Etre capable de surveiller son environnement et d'anticiper ses évolutions, quelle que soit la langue dans laquelle les informations sont exprimées, est une démarche vitale aux entreprises pour maintenir ou développer leur compétitivité. L'information est au cœur d'une telle démarche d'Intelligence Economique. En effet, un grand nombre de documents publics et disponibles sur Internet (dépêches de presse, bases de données bibliographiques, scientifiques et techniques,) ou en Intranet (mails électroniques, rapports techniques) contiennent potentiellement de l'information utile à la décision. L'enjeu est donc de pouvoir analyser, classifier et organiser des documents textuels, afin de permettre à des utilisateurs non-experts de fouiller ces documents et/ou de les évaluer.

Notre article est une illustration des technologies de text mining développées au sein de la société TEMIS, et plus précisément celles d'extraction d'information. Contraction de Text Mining Solutions, TEMIS est une société qui conçoit et propose des applications dédiées à l'analyse textuelle, sur de gros volumes de données dites non structurées.. Le serveur d'extraction d'information, TEMIS Insight Extractor™, couplé à des Skill Cartridge™ thématiques (intelligence économique), ou spécialisées par domaine d'activité (industrie pharmaceutique) propose des applications dédiées à la veille stratégique et concurrentielle, à la gestion de la relation clients, à la gestion de la connaissance et des savoir-faire et à la gestion des ressources humaines.

Après un court rappel sur l'extraction d'information, nous décrirons, dans un premier temps, notre approche sur le formalisme et la méthodologie de construction de règles d'extraction d'information par niveaux implémentés dans les Skill Cartridge™ couplées au serveur d'extraction d'information. Notre exposé sera illustré d'exemples concrets issus de deux applications réelles d'intelligence économique mis en œuvre conjointement avec les équipes des deux entités mentionnées ci-dessous.

2 L'extraction d'information

Selon Cowie et Wilks, l'extraction d'information (*Information Extraction, IE)"is the name given to any process which selectively structures and combines data which is found, explicitly stated or implied, in texts"* [Wilks, 1997]. En d'autres termes, l'extraction d'information est le processus qui permet d'identifier l'information pertinente, les critères de pertinence étant exprimés sous forme de patrons d'extraction *(extraction patterns)* et les données extraites transposées dans des formulaires *(templates*) prêts à être remplis.

2.1 L'approche TEMIS

L'un des enjeux essentiels de notre activité est la capacité à développer une application opérationnelle pouvant s'adapter aisément à de nouveaux domaines d'application, de nouveaux secteurs d'activité, de nouvelles thématiques d'analyse ou à une autre langue. Notre démarche s'est naturellement orientée vers la validation, sur une base empirique, de la cohérence de formalismes existants [Grishman, 1997], [Neumann, 1999]. Les critères retenus étant :

- la facilité d'implémentation
- la maintenabilité
- la réutilisabilité

Pour intégrer aisément des données spécifiques (sociétés ou acteurs du domaine, produits du domaine, actions propres au domaine,...) à une thématique d'analyse (intelligence économique, analyse de courriels, étude de CV), nous avons donc implémenté le formalisme choisi et défini une méthode privilégiant les aspects multi-domaines et/ou multilingue. L'objectif, à terme, est d'améliorer les conditions d'utilisation du système de manière à ce que l'effort nécessaire pour sa personnalisation par des experts (capitaliser/valoriser les connaissances de leur domaine d'application), soit 'raisonnable'. On observe, en effet, que le coût de la personnalisation des systèmes d'extraction est particulièrement élevé si les modifications doivent être effectuées par les développeurs des règles d'extraction (le plus souvent des linguistes).

2.1.1 Le formalisme

L'objectif est de construire des patrons d'extraction (extraction patterns) [Yangarber et Grishman, 1997] suivant une approche guidée par le but [Appelt, 1993] [Poibeau, 2002]. Un patron d'extraction décrit une structure syntaxique de surface comportant des éléments lexicaux et/ou amorces (trigger words), des tags grammaticaux et des éléments typés sémantiquement. En d'autres termes, un patron d'extraction est une expression régulière qui identifie le contexte de syntagmes pertinents et les délimiteurs de ces syntagmes.
Les règles d'extraction sont exprimées sous forme d'expressions régulières combinant l'accès aux formes de surface, aux tags grammaticaux et aux lemmes. En associant un concept à un patron, une règle ajoute de l'information à une séquence de mots, par exemple, en lui attribuant un nom de classe sémantique qui peut être ensuite utilisé dans d'autres règles. Le module d'extraction utilise la technologie des transducteurs [Hobbs 1997].

2.1.2 Skill Cartridge™ et méthodologie

2.1.2.1 Structure modulaire

Notre objectif étant d'assurer maintenabilité et réutilisabilité, nous modélisons et organisons l'information à extraire selon une hiérarchie de composants de connaissance modulaires intégrables à différents domaines d'activité et/ou langues. Cette hiérarchie est appelée Skill Cartridge™ ou cartouche de connaissance. Un composant de connaissance peut avoir la forme d'un ou de plusieurs dictionnaire(s) ou d'un ensemble de règles d'extraction.
Notre approche de développement des composants de connaissance est guidée par l'idée de favoriser leur réutilisation, de la même manière que dans un langage de programmation, il est possible de définir des classes d'objets et de les utiliser. La méthodologie développée concerne plusieurs tâches :

- La définition de concepts outils ou macros de 1er niveau
- la décomposition de la construction des patrons d'extraction : par langue au niveau de la définition des concepts de base, puis générique au niveau des concepts supérieurs, et des règles de liaison des différents patrons.
- l'organisation hiérarchique des patrons d'extraction en niveau d'extraction.
- l'organisation des fichiers sous forme d'arborescence. Pour un domaine spécifique, il est possible d'étendre ou de redéfinir un concept descripteur avec du lexique ou des règles sans modifier le corps de la cartouche.

2.1.2.2 Hiérarchisation de l'information en niveaux d'extraction

Notre système utilise la règle classique « le plus à gauche, le plus long » pour savoir quelle séquence de mots associer aux patrons candidats. Afin de gérer des priorités différentes, une cartouche peut être décomposée en niveaux contenant chacun un sous-ensemble d'expressions. Un concept extrait à un niveau donné encapsule les unités qui l'ont déclenché, rendant celles-ci inaccessibles aux niveaux supérieurs et permettant d'utiliser ce concept pour en bâtir d'autres.
La hiérarchisation de l'information par niveaux permet ainsi de gérer des hiérarchies de patrons et de contrôler leur exécution sur les corpus. Toute séquence de mots regroupée au sein d'un patron, l'est de façon définitive pour tous les niveaux suivants et ne pourra donc pas être disloquée pour construire un patron concurrent à un niveau supérieur. Un mot isolé, qui n'a pas participé à la construction d'un patron reste disponible pour la construction d'un autre patron en changeant de niveau.

2.2 Les solutions proposées

Cette approche privilégiant à la fois l'aspect modulaire des composants de connaissance et l'organisation hiérarchique de ceux-ci a permis de développer des applications adaptées aux demandes de nos clients, pour lesquelles les critères retenus ont été la pertinence des extractions et la flexibilité de la solution proposée. Deux exemples d'applications réelles, liée à l'intelligence économique sont présentés, ci-dessous, en illustration de nos propos.

3 Veille concurrentielle : génération automatique de fiches entreprises

La première application a été mise en œuvre pour le groupe France Télécom dans le cadre du projet Extractor. Au sein du service de veille concurrentielle, marketing et financière du groupe, l'équipe TEMIS a mis en œuvre un processus de création et de mise à jour semi-automatique de monographies d'entreprises à partir d'un flux de dépêches Reuters.

Le projet Extractor, préparé et mis au point avec la société TEMIS, a été cofinancé par deux unités d'affaires de France Télécom, l'ARIA (Agence en Réseau pour l'Information Active) et la DBI (Direction de la Business Intelligence). Ces deux unités d'affaires proposent des services et des produits de veille élaborés, la première pour le groupe France Télécom, la seconde en externalisant ces services.

3.1 Le projet

3.1.1 La problématique

Au sein de France Télécom, nombreuses sont les unités d'affaire à mettre à disposition, sur leur site intranet, des *fiches entreprises*, ou *monographies*. Cependant, au regard de la très rapide évolution du secteur, la mise à jour *manuelle* de telles informations est purement impossible. Le flux journalier de dépêches reçues sur le serveur de l'Aria est très volumineux (3 000 à 4 000). Malgré une catégorisation en amont proposée par des agences de presse comme Reuters, le tri des informations sur une société donnée ne peut être totalement fiable, et ne répond pas expressément aux besoins propres de l'organisation. Le processus de création et de mise à jour semi-automatique de fiches entreprises revêt donc un intérêt stratégique pour une bonne connaissance des acteurs du secteur des télécommunications, de leurs produits et de leur activité.

3.1.1.1 Les objectifs du projet

Le processus mis en œuvre au sein du projet *Extractor* consiste à explorer une base documentaire, en extraire les informations pertinentes et remplir un formulaire préalablement défini. Ce formulaire constitue le corps de la monographie et regroupe les différentes rubriques qui doivent être renseignées pour former la fiche entreprise.

Les flux de presse, rédigés en langue anglaise, ont été pendant la durée du projet analysés par **Insight Discoverer™ Extractor** couplé à une **Skill Cartridge™** dédiée **Intelligence Economique** (en anglais). En guise de résultat**,** les informations extraites alimentaient une fiche entreprise.

3.1.2 La monographie Extractor

3.1.2.1 Le fond…

Les informations de la fiche ont été définies avec l'aide d'un expert, en différenciant les informations devant être renseignées manuellement, et ce, de façon quasi-définitive (nom, raison sociale, numéros de téléphone…), des informations susceptibles d'évoluer. Ce sont ces dernières qui doivent être extraites automatiquement, à partir d'un flux de dépêches.

Les rubriques

On présente, ci-dessous, les rubriques à renseigner dans une monographie type :

CADRE JURIDIQUE
GOUVERNANCE
DIRIGEANTS
EFFECTIFS
DONNEES FINANCIERES
FILIALES-PARTICIPATIONS-ACQUISITIONS
PARTS DE MARCHE
NOMBRE DE CLIENTS
QUESTIONS COMMERCIALES

PARTENARIATS
FOURNISSEURS
CONTENTIEUX
DESCRIPTIF RESEAU

Le pilote réalisé pour évaluer la viabilité de ce projet a consisté à mettre en place un processus de création et de mise à jour d'une fiche d'un acteur des télécommunications : France Télécom. Il était ainsi aisé de valider ou d'invalider les informations extraites.

3.1.2.2 … et la forme

En termes de produit fini, la forme est un facteur tout aussi important que le fond. Lorsque l'information est extraite, l'utilisateur final doit avoir la possibilité de retrouver le contexte dans lequel elle est apparue. Ainsi, pour chaque information présente dans la fiche doivent être mentionnés (Cf. fig. 2, colonne droite) :

- le titre de la dépêche correspondant à l'extraction ;
- la date de la dépêche ;
- la phrase entière concernée par l'extraction ;
- le lien hypertexte permettant d'accéder à la dépêche originale.

Ces informations servent à restituer le contexte de l'extraction. Il est possible de retourner rapidement au document initial, la phrase concernée par l'extraction permettant de lever toute ambiguïté. Il est également possible de mettre en place des filtres sur la date ou la source pour ne visualiser que des extractions portant sur un type de document précis ou sur des documents émis à partir d'une date donnée.

Le formulaire

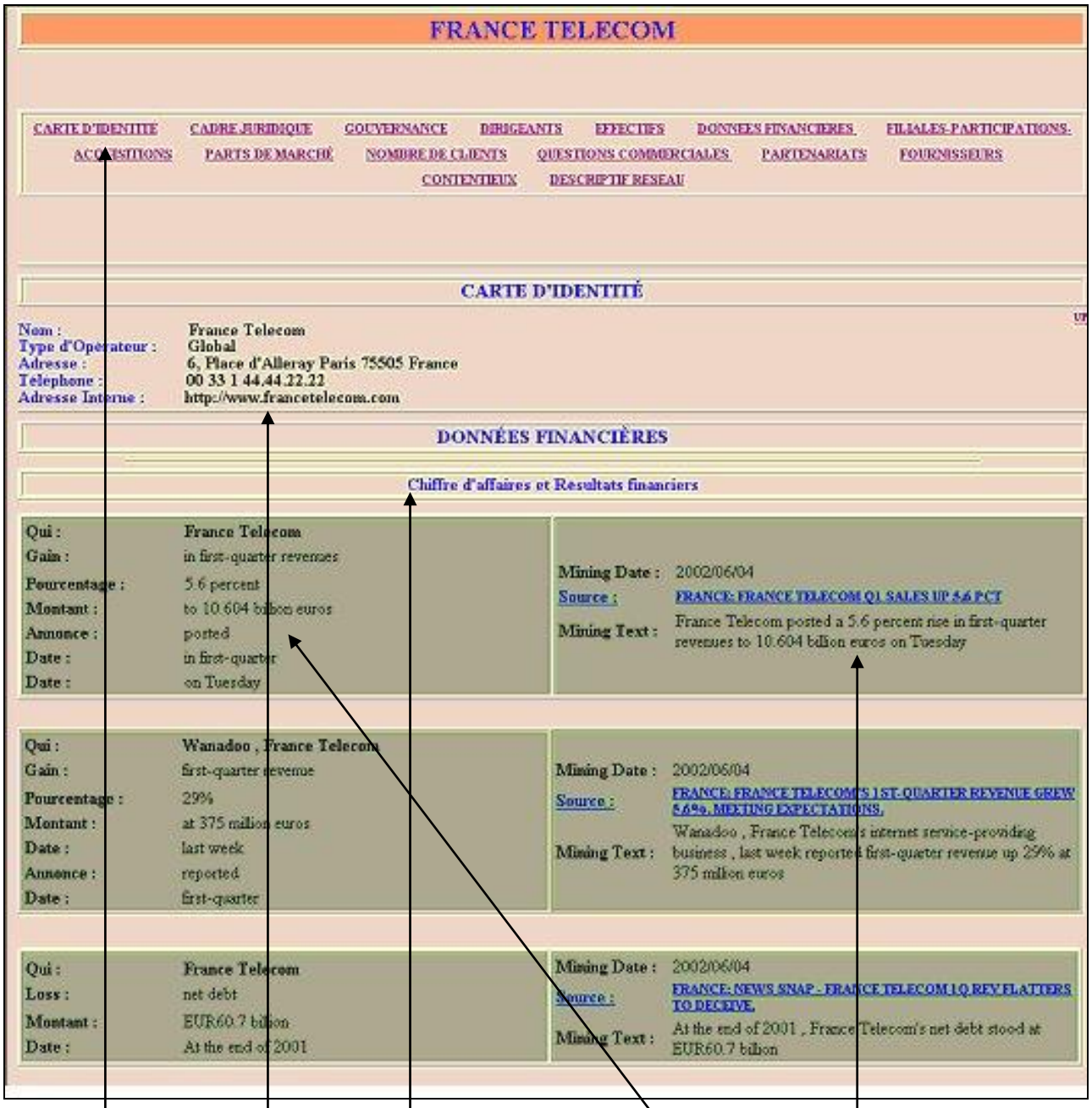
Les extractions effectuées vont remplir un formulaire dont le cadre est fourni ci-dessous. Les zones définies se présentent sous forme d'attributs valeurs.
La colonne de gauche traite de l'extraction proprement dite. En fonction des informations présentes dans la phrase extraite, elle précise quel est l'acteur (Qui), le thème de l'extraction (annonce d'un résultat financier par exemple), les informations chiffrées s'y rapportant (pourcentage, montant), et la date de l'événement.

Qui **Finance** **Pourcentage** **Montant** **Date**	**Mining Date :** Date **Source :** **Titre de la dépêche** **Mining Text :** Phrase source

Figure 2 : **Exemple de formulaire**

3.1.3 La réalisation du pilote

A partir des données extraites, un filtre portant sur un acteur des télécommunications permet de regrouper les informations le concernant dans une fiche dédiée. La figure ci-dessous en présente un extrait.



Rubriques de la fiche

Attributs valeurs renseignés

Rubrique Carte d'Identité

Information relative à l'extraction et au document concerné

Sous-rubrique Chiffre d'affaires et Résultats financiers

3.2 L'évaluation du projet

3.2.1 Le jeu de tests

Durant le pilote, deux jeux de dépêches ont été constitués. Le premier a servi de base aux tests d'extraction pour mesurer et améliorer les valeurs de précision et de rappel. Ce jeu de dépêches a été pleinement exploité pour perfectionner et affiner les règles d'extraction. Le second, constitué au moment des tests finaux, a permis d'évaluer la qualité des règles d'extraction.

3.2.1.1 La customisation des règles d'extraction

La qualité des outils de text mining se mesure par le rappel et la précision :

$$rappel = \frac{nombre\,d'extractions\ pertinentes\ obtenues}{nombre\,d'informations\ pertinentes\ dans\ la\ base}$$

$$précision = \frac{nombre\,d'extractions\ pertinentes\ obtenues}{nombre\,d'extractions\ obtenues}$$

Idéalement, le rappel et la précision devraient tendre vers 100%. Un *rappel* de 100% signifie que toutes les informations pertinentes dans la base ont été extraites, le système n'ayant *omis* aucune d'entre elles. Une précision de *100%* signifie, quant à elle, que toutes les extractions obtenues sont pertinentes ; il n'y a pas de bruit.

3.2.1.2 Les résultats de l'amélioration des règles d'extraction

Le premier jeu de 300 dépêches, pour la première fois testé en décembre 2001, obtenait des taux de rappel et de précision inférieurs à 50%. Ce corpus a ensuite été exploité afin d'améliorer les règles d'extraction et de parvenir à de meilleurs résultats.

Un partenariat avec l'université de Poitiers a permis d'affecter des étudiants à la tâche d'analyse des résultats, dans le processus d'amélioration de la qualité des règles et des dictionnaires métier. A l'issue du second test en avril 2002, les résultats se sont améliorés de 30 points.

Le risque était que ces scores soient dus à la *surexploitation* du premier corpus, et que le système montre de piètres performances sur un corpus *vierge* de toute exploration. Or, les résultats, en termes de rappel pour le moins, sont aussi bons, (voire meilleurs en fonction des rubriques !) sur le nouveau corpus de 311 dépêches.

Figure 12 : **Tableau des résultats**

		Nbre Phrases du fichier	Doublons	Nbre Phrases à ne pas traiter	Nbre Phrases à extraire	Phrases Complexes	run_1 dated 02	run_2 dated 03	run_3 dated 04/09
Total		374	60	45	**269**	43	102	150	**191**
							38,50%	56%	**71%**

Ce tableau retrace l'évolution de la qualité des extractions en termes de réduction du silence : phrases qui devraient être extraites, mais qui ne le sont pas.

A partir du corpus initial de 300 dépêches, un large panel de phrases (374) à extraire a été constitué manuellement. Chacune de ces phrases a été traitée par l'extracteur, et les résultats ont été systématiquement analysés afin d'améliorer les règles d'extraction. Ainsi, sur le total des 374 phrases, 102 phrases (38,5%) étaient extraites en avril. En septembre, ce nombre est passé à 191, soit plus de 70%.

4 Veille stratégique : émission de rapports quotidiens

La seconde application concerne un projet de prospection d'implantations d'entreprises étrangères développé au sein d'un organisme de développement économique national.

Cet organisme a pour objectif de développer des implantations durables d'entreprises étrangères, créatrices d'emplois et de richesse. Pour remplir au mieux ces missions, la Direction des Etudes Stratégiques (DES) a développé des techniques d'intelligence économique, à partir des outils TEMIS, pour orienter sa stratégie, cibler ses opérations de prospection et enrichir ses connaissances.

4.1 Le projet

4.1.1 Le contexte

La cellule de veille mise en place par la DES a deux principales missions : suivre les évolutions du marché des investissements internationaux et aider les correspondants en poste en France et à l'étranger à identifier leurs prospects.

Afin de connaître précisément le marché des investissements internationaux et ses évolutions la DES a créé et développé deux observatoires économiques ; l'un recense les investissements en France, l'autre les investissements internationaux en Europe. Les informations de ces observatoires enrichissent des bases de données qui sont utilisées pour produire des bilans et rapports en temps réel.

Pour d'améliorer la pertinence de ses travaux de prospection, la DES a également entrepris d'identifier les entreprises ayant des projets d'investissement. Ces informations alimentent une base de données qui permet l'émission de fiches destinées à orienter les démarches de prospection des correspondants en France et à l'étranger.

4.1.2 La problématique

Ces trois bases de données sont alimentées par les analystes chargés de la veille, après validation des informations extraites. Le développement du système de traitement de l'information a commencé par une phase importante d'identification des sources pertinentes et de paramétrage du système de crawl.

4.1.2.1 Une sélection trans-sectorielle et internationale

Notre client s'intéresse aux implantations à l'échelle internationale et ce, à travers tous les secteurs d'activités. Avec un spectre aussi large, nombreuses sont les sources d'information disponibles. Un long travail d'identification et de qualification des sources a été effectué. Les outils de rapatriement de l'information ont ensuite été paramétrés afin de définir les termes clé utilisés pour attaquer les moteurs de recherches et autres bases de données.
Le vocabulaire utilisé pour la recherche et le rapatriement de l'information est assez générique, il s'agit de mots clé tels que « implantation, nouvelle usine, transférer, délocaliser » etc., de telles recherches génèrent nécessairement beaucoup de bruit (information non pertinente), l'objectif premier étant de ne pas rater d'information pertinente (éviter le silence).

4.1.2.2 Une information croissante

Ainsi, à mesure que la couverture des sources atteignait un niveau satisfaisant, les analystes se trouvaient face à une masse d'information quotidienne de plus en plus difficile à exploiter.
Les différents systèmes de crawl et surveillance de page mis en place ont rapidement rendu la masse de données rapatriée impossible à exploiter.

Exemples :
*1) Ils étaient nombreux hier matin à occuper l'école de Pujaudran pour refuser cette **délocalisation** de 12 jeunes Pujaudranais alors que d'autres solutions peuvent être envisagées.*
*2) Samedi, au Mazet-Saint-Voy, une météo tristounette a accompagnée l'inauguration des **nouvelles installations** du stade et du camping municipal.*
*3) Parallèlement, les enquêteurs soupçonnent Charles Pieri d'avoir participé au **transfert** de certains joueurs et de s'être comporté comme un dirigeant de fait du club de Bastia.*

*4) Le premier groupe russe d'aluminium, Roussal, a annoncé lundi le lancement d'une **nouvelle usine** de fabrication de canettes en aluminium à Vsevolojsk, près de Saint-Pétersbourg (nord-ouest).*
*5) Air Liquide a **délocalisé** son usine d'acétylène installée à Saint-Yrieix.*

Pour distinguer les phrases bruyantes des informations pertinentes, il fallait considérer la syntaxe et bâtir des modèles de phrases pertinentes.
La solution apportée devait remplir deux objectifs :

- filtrer les informations pertinentes parmi tous les documents rapatriés par le système
- présenter les résultats sous une interface unique afin de limiter le temps passé à manipuler les différentes sources.

4.1.3 La réponse apportée par l'extraction d'information

Le système de text mining mis en œuvre consiste à extraire de la masse de données rapatriée chaque jour par le crawler, une information pertinente et qualifiée.
Une Skill Cartridge™ spécifique a été développée pour répondre aux besoins du client. Les informations stratégiques ont été listées avec les analystes et modélisées par TEMIS dans la Skill Cartridge™ dédiée (à ce jour en français et en anglais).

Les informations extraites concernent trois problématiques :

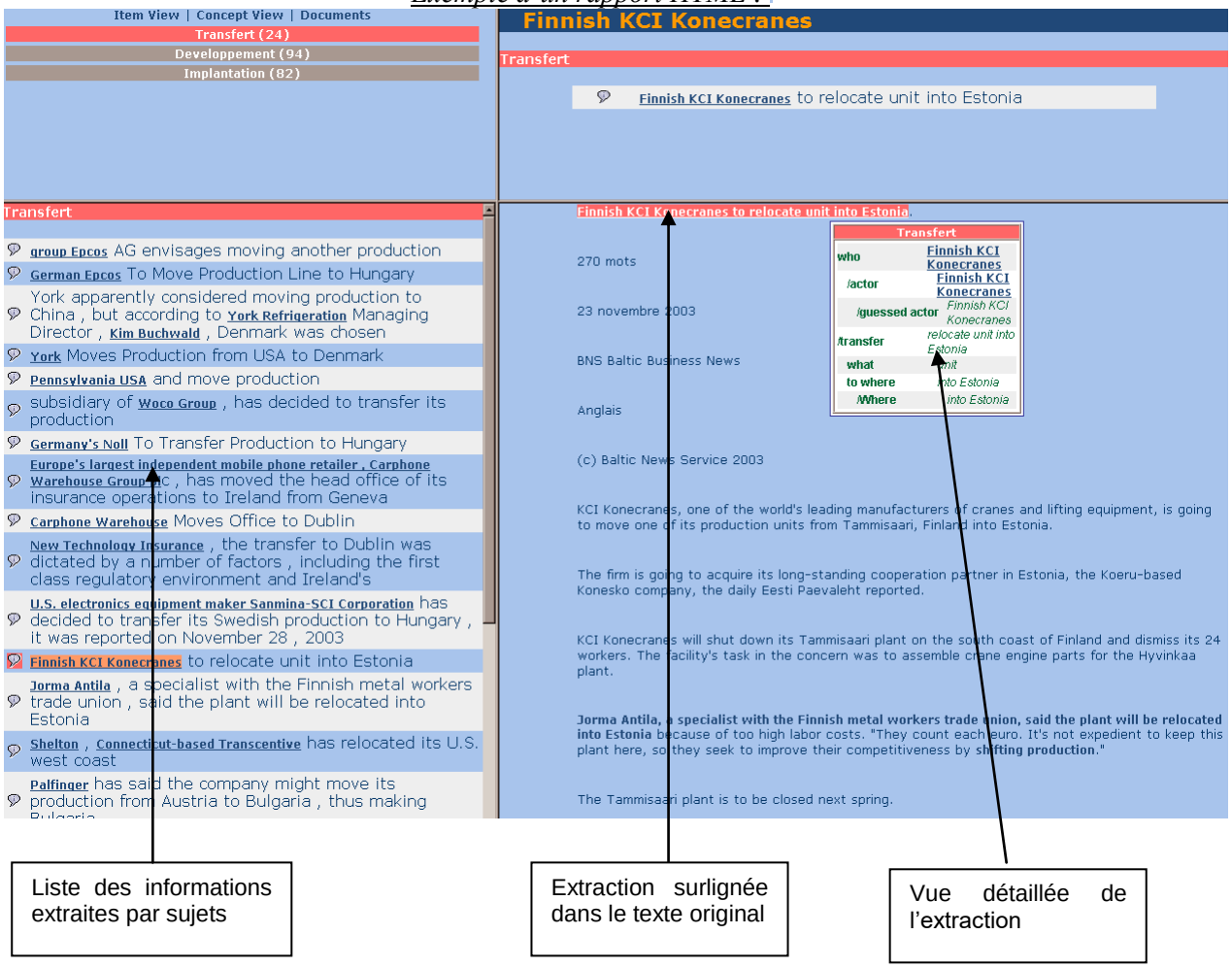
- Les implantations : ouvertures de sites, constructions d'usines, ouvertures de filiales, de nouveaux bureaux.
- Les transferts : transferts d'usine, de production.
- Les développements : expansion d'une société dans une zone géographique, consolidation de sa présence quelque part, et les investissements.

L'ensemble des sources d'information sont donc quotidiennement analysées par Insight Discoverer™ Extractor couplé à une Skill Cartridge™ dédiée Investissements Internationaux (en français et en anglais). Le système produit chaque matin un rapport HTML présentant les informations stratégiques pour les analystes.

4.1.3.1 Rapport HTML quotidien

Pour faciliter la lecture des documents le rapport HTML publié chaque matin présente l'information extraite dans trois principales catégories : Les implantations, les transferts d'activité et les développements (développements géographique, investissements…).
Un clic sur une des trois catégories dans la fenêtre en haut à gauche ouvre une fenêtre située en bas à gauche dans laquelle apparaissent les extractions. La lecture de cette fenêtre permet à l'analyste de juger de la pertinence de l'information et de choisir ou non d'ouvrir l'article correspondant par un clic sur l'icône.
En cliquant sur l'entreprise citée dans la fenêtre de gauche on obtient, en haut à droite toutes les extractions relatives à cette entreprise.

Exemple d'un rapport HTML :



4.1.3.2 Modélisation des règles d'extraction

La problématique de cet organisme de développement économique national se prête parfaitement à l'extraction car si les termes permettant de détecter l'information constituent un vocabulaire assez générique, l'utilisation de la syntaxe permet de lever rapidement la grande majorité des ambiguïtés.
Les règles d'extraction ont été modélisées sur la base d'un corpus de référence constitué par les analystes présentant un ensemble de phrases pertinentes en français et en anglais. Un autre corpus a été constitué rassemblant des phrases bruyantes, c'est-à-dire des phrases dans lesquelles un terme clé déclencheur est présent, ce qui a entraîné le rapatriement du document, mais celui-ci n'est pas pertinent pour notre problématique (cf. exemple 1, 2 et 3).
L'objectif était de bâtir les patrons d'extractions pour les phrases que l'on cherchait à extraire tout en vérifiant que les phrases bruyantes, elles ne l'étaient pas.

4.2 Les perspectives

4.2.1 Extension de la Skill Cartridge™ existante

Une des raisons du choix de la solution proposée par TEMIS a été la possibilité pour les analystes de la cellule de veille d'étendre eux-mêmes la couverture de la Skill Cartridge™ Investissements Internationaux. TEMIS les accompagne dans la définition des priorités et assure un transfert de compétence pour la mise à jour des concepts existants et la construction de nouveaux concepts.

4.2.1.1 Ajout de nouvelles langues

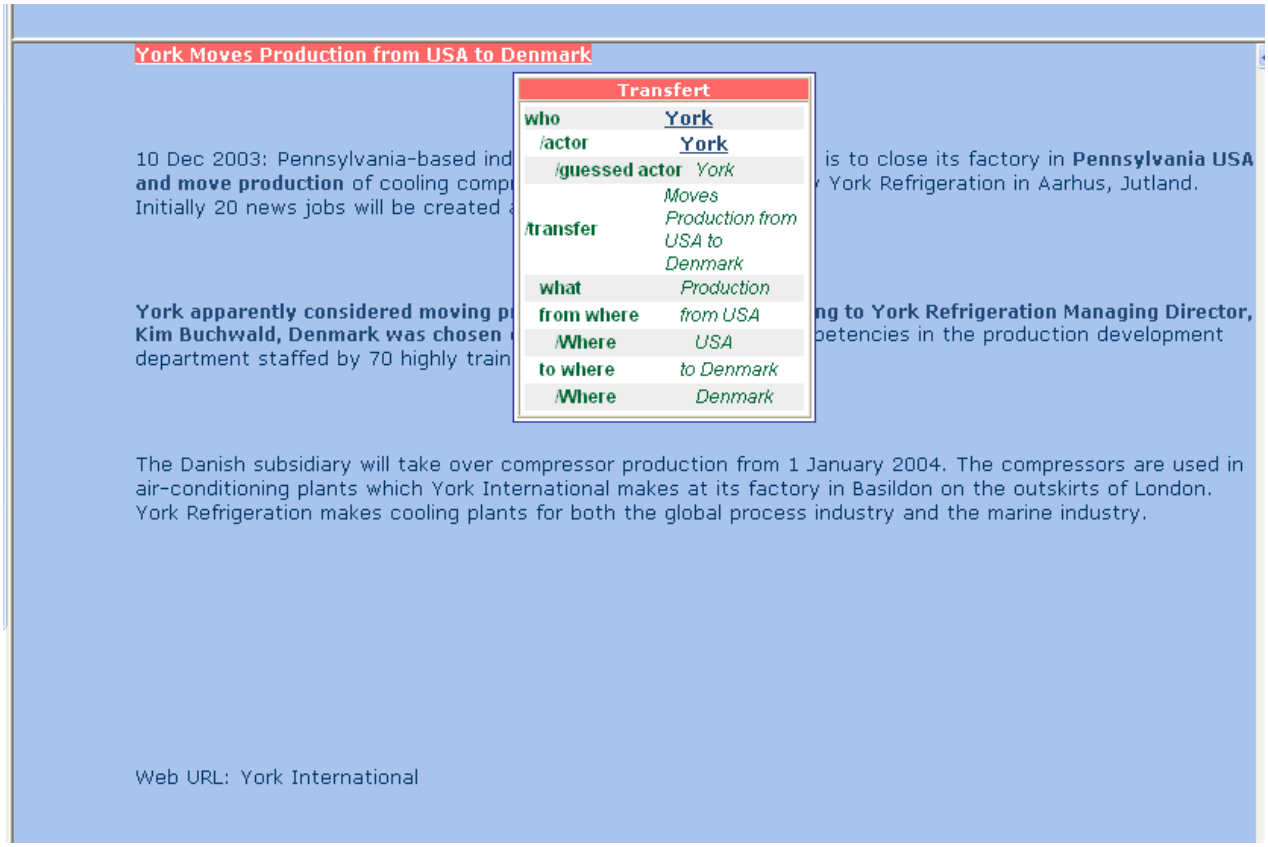
L'architecture de la Skill Cartridge™ facilite le portage vers d'autres langues. Les règles d'extractions sont partagées par différentes langues, en créant les dictionnaires dans le même modèle que les langues existantes on obtient une base exploitable que l'on peut ensuite faire évoluer en fonction des spécificités de la nouvelle langue.

Les analystes travaillent actuellement au développement des concepts en langue espagnole.

4.2.1.2 Ajout de filtres géographiques

Une des améliorations envisagées prochainement concerne le domaine des délocalisations et les transferts d'activités. La hiérarchie des dictionnaires des noms de lieu et les contraintes de syntaxe permettent de distinguer la destination d'une délocalisation de son point de départ et de limiter ainsi l'affichage des extractions aux cas pertinents pour la problématique spécifique de notre client.

Dans l'exemple ci-dessous on distingue le « from where » « from USA » du « to where » « to Denmark ». Cette délocalisation des Etats-Unis vers l'Europe est particulièrement pertinente dans le cadre de ce projet et sera à terme qualifiée comme telle.



5 Conclusion

Cet article met l'accent sur les capacités des technologies d'extraction d'information à répondre à des besoins différents, et à développer des solutions adaptées à ces besoins. En effet, en se fondant sur la même technologie, TEMIS a mis en œuvre pour deux organismes totalement différents, deux solutions de veille diversifiées, mais répondant parfaitement aux besoins exprimés par chacun d'eux.
Les atouts de cette technologie se mesurent en termes de qualité des extractions et de flexibilité des solutions proposées. La stratégie de TEMIS est de développer des composants de text mining modulaires pouvant facilement s'intégrer dans toute application client visant à extraire l'information pertinente à partir d'une collection de documents.

Au niveau fonctionnel, les points forts des solutions proposées par TEMIS sont de permettre une automatisation de la collecte d'information et de s'adapter aux données à analyser. Surtout, la technologie d'extraction mise en œuvre permet de prendre en compte le profil de l'utilisateur et la diversité des objectifs de l'extraction, qu'il s'agisse de problématiques de veille, de gestion de la relation clients ou de gestion de la connaissance.

Au niveau opérationnel, le text mining trouve des applications critiques dans le domaine de la Santé et plus particulièrement pour l'industrie pharmaceutique. Confrontée à la croissance exponentielle de l'information textuelle depuis la complétude du séquençage du génome humain, l'industrie pharmaceutique est aujourd'hui en quête de facteurs lui permettant d'accélérer son processus de recherche fondamentale, de détecter au plus tôt les potentiels effets secondaires de molécules en développement ou encore, de positionner au mieux son futur médicament face à la concurrence, etc. Dans une industrie où, des publications scientifiques aux commentaires des médecins examinateurs ou aux importants fils de presse en temps réel, l'information est avant tout textuelle, le text mining constitue un apport majeur.

Références

[1] APPELT D., HOBBS J., BEAR J., ISRAEL D., KAMEYAMA M. ET TYSON M. *FASTUS : a finite-state processor for information extraction from real-world text.* In proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'93), Chambéry, 1993, pp. 1172-1178.

[2] AUBRY C, GRIVEL L, GUILLEMIN-LANNE S, LAUTIER C *Aide à la construction de composants de connaissance pour l'extraction d'information : méthodologie et environnement* , CIFT 2002, Hammamet-Tunisie, 21-23 octobre 2002.

[3] BUSHBECK B, GRIVEL L, GUILLEMIN-LANNE S, LAUTIER C *Une application industrielle d'extraction d'informations pour l'Intelligence Economique*, EGC 2002 Extraction et Gestion des Connaissance, Montpellier, 21-23 janvier 2002.

[4] EPPSTEIN R., *Création d'un système d'information stratégique dans le domaine des technologies de l'information et de la communication – Application à CS Communication & Systèmes*, Thèse, Université de Marne-la-Vallée, 28 novembre 2001.

[5] GRISHMAN R. *Information Extraction: Techniques and Challenges.* In M.T. PAZIENZA (éd.), Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, Springer Verlag, Heildelberg, 1997, pp. 10-27.

[6] GRIVEL L, GUILLEMIN-LANNE S, COUPET P, HUOT C *Analyse en ligne de l'information: une approche permettant l'extraction d'informations stratégiques basée sur la construction de composants de connaissance*, VSST 2001, Barcelone, 15-19 octobre 2001

[7] HOBBS J. R. ET AL. FASTUS, *A Cascaded Finite-State Transducers for Extracting Information from Natural-Language Text.* In E. Roche et Y. Schabes (eds.), Finite-State Language Processing. Cambridge MA: MIT Press, 1997

[8] NEUMANN G., SCHMEIER S., *Combining Shallow Text Processing and Machine Learning in Real World Applications*, Proceedings of the IJCAI-99 workshop on Machine Learning for Information Filtering, Stockholm, Sweden, 1999.

[9] POIBEAU T. *Extraction d'information à base de connaissances hybrides*, Thèse, Université Paris-Nord, 8 mars 2002.

[10] WILKS, Y.*Information Extraction as a Core Language Technology*. In Pazienza, M.T., editor, Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, Frascati, Italy, LNAI Tutorial, Springer. pp. 14-18,1997.

[11] YANGARBER R., GRISHMAN R., *Customisation of Information Extraction Systems*. In Pazienza, M.T., editor, Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, Springer Verlag, Heildelberg, 1997, pp. 1-11.

[12] ZANASI, A. *Text Mining: The New Competitive Intelligence Frontier. Real Application Cases in Industrial, Banking and Telecom/SMEs World*, VSST 2001, Barcelone, 15-19 octobre 2001.