MAITRISER LE PROCESSUS DE TEXT MINING DANS LE CADRE D'APPLICATIONS D'INTELLIGENCE ECONOMIQUE, DE GESTION DE LA RELATION CLIENT OU DE GESTION DE CONNAISSANCES

Luc Grivel (*, **)
Luc.Grivel@univ-paris1.fr

(*)Chercheur associé, Equipe INGENIERIE DES SYSTEMES D'INFORMATIONS STRATEGIQUES ET DECISIONNELLES,
Université de Marne-La-Vallée, France
(**)Maître de Conférences à l'Université de Paris1, 17 rue de la Sorbonne, 75 005 Paris France

Mots clefs :
Fouille de données textuelles, analyse de textes, Intelligence économique, Intelligence client, gestion des connaissances, ingénierie des connaissances,
Keywords:
Text mining, text analytics, competitive intelligence, knowledge management, management, knowledge engineering.
Palabras clave :
administración del conocimiento, ingeniería del conocimiento,

Résumé

L'article décrit un processus de text mining combinant extraction d'information, clustering et classification supervisée pour analyser des enquêtes de satisfaction chez un constructeur automobile. L'objectif est de collecter, analyser plus efficacement les retours-clients sur les nouveaux modèle pour déterminer le plus rapidement possible d'où proviennent les dysfonctionnements ou problèmes mentionnés par les clients (quelles sont les parties de la chaîne de production ou de la chaîne de distribution impliquées dans le problème).

Après avoir passé en revue les principales technologies de text mining, l'auteur tire les leçons de ce projet pour illustrer le fait que, au-delà des performances des outils de text mining, il faut, pour transformer l'information brute en information utile, exploitable, imaginer des processus de traitement adaptés aux réalités du terrain (données, bruitées, manquantes, temps disponible, …). Il dresse ensuite un rapide panorama des principales applications pouvant tirer parti des techniques de text mining : gestion de la relation client, intelligence économique et gestion des connaissances.
.

1 Introduction

La mondialisation de l'économie conduit les acteurs (états, organisations, compagnies, citoyens, employés, clients), à porter une attention accrue aux relations qu'ils entretiennent, le plus souvent dans une optique d'anticipation des menaces ou de détection des opportunités de développement. Plus précisément, si on se limite à l'entreprise, la matière première d'une analyse de toutes ses relations avec les acteurs de son environnement est constituée des documents qu'elle échange, avec ses clients, partenaires, employés, fournisseurs, administrations, etc. Aujourd'hui, cette matière première est abondamment disponible et accessible en ligne sous forme électronique sur les réseaux informatiques internes (Intranet) ou externes.

Comment repérer ces relations, mieux exploiter les informations stratégiques enfouies dans ces documents ? Lorsque les volumes de documents en jeu sont tels qu'une analyse manuelle de leur contenu n'est plus possible, un ensemble de technologies, désignées sous le nom de text mining, disponibles sous forme de composants logiciels, permettent de simplifier la recherche, l'organisation et l'analyse des contenus numériques entrant et sortant de l'entreprise. Pour autant, les réalités du terrain (données bruitées, données manquantes, ressources humaines et outils disponibles, …) demande d' imaginer des processus de traitement adaptés, bref, de développer une ingénierie des connaissances.

Après avoir passé en revue les principales technologies de text mining, en illustrant, par quelques exemples, le type d'information qu'elles permettent de traiter, nous prenons l'exemple d'un projet réalisé chez un constructeur automobile pour illustrer le processus de text mining qui a été mis en place pour traiter des enquêtes de satisfaction puis nous dressons un rapide panorama des principales applications pouvant tirer parti des techniques de text mining : gestion de la relation client, intelligence économique et gestion des connaissances.

2 Technologies pour analyser et organiser les contenus

Les techniques permettant de mener à bien une analyse automatique du contenu de corpus de documents textuels numériques sont connues sous le nom de text mining. Elles combinent des approches, issues de disciplines comme la linguistique informatique (analyse de texte et extraction d'information), l'analyse de données (analyse factorielle, classification, …), l'intelligence artificielle (techniques d'apprentissage, règles d'inférences) et les sciences de l'information (lois bibliométriques). Nous les décrivons ici sommairement sur un plan fonctionnel. Le lecteur pourra se référer à [3,4] pour une description plus technique.

2.1 Analyse de texte et extraction d'information dans un corpus multi-lingue

Selon 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"* [10]. Les informations que l'on cherche à repérer peuvent se trouver sous des formes ou des langues diverses [1].
« Avec une participation de 29.0%dans un important champ pétrolier (Buzzard Oilfield) situé en Mer du Nord, Petrocanada …. »

« Petrocanada has acquired a 29.9% of interest in the Buzzard North Sea oilfield in 2003 »

Dans ces deux extraits, l'objectif est d'identifier le prédicat **Prise de participation** ainsi que ses arguments.

Prise de participation : acquisition/acquire
Acquéreur : Compagnie/Petrocanada
Cible : Champ pétrolier/Buzzard Oilfield
Montant de part : 29.9%
Date : Année/2003
…

Figure 1 : extraction d'information sur des dépêches de presse[1]

Dans une note transcrite au niveau d'un centre d'appel, l'objectif pourra être d'identifier des opinions négatives d'un client et ses intentions
Cstmr XXX not happy with his cell phone – customer wants to switch to Orange

Client : XXX
Opinion negative : not happy
Produit : cell phone
Client : XXX
Intention : wants
Action negative : switch to
Concurrent : compagnie de téléphone/Orange

Figure 2 : extraction d'information sur des réclamations d'un client[2]

Sur le plan technique [3,4], un système d'extraction d'information effectue sur une phrase 4 analyses successives:

- lexicale (segmentation de la phrase en chaînes de caractères qui représenteront des mots) et identification de la langue,
- morpho-syntaxique (étiquetage des mots par leur catégorie syntaxique et association de chaque mot à sa forme canonique ou pseudo-racine : acquired → VerbeParticipePassé Acquire),

1 Anne-Geneviève Bonnet-Ligeon (Total) lors d'un exposé à la Sorbonne le 21 mars 2005.
2 Olivier Jouve (SPSS) lors d'une table ronde aux 1ères rencontres Innovation, Compétitivité et Connaissances, le 28 septembre 2005

- syntaxique (analyse de la structure de la phrase, en se limitant à reconnaître mot par mot les frontières majeures de constituants (syntagmes nominaux ou verbaux) et des relations entre les mots, à partir de règles de grammaires locales,
- sémantique (compréhension du sens des mots et des relations entre les mots).

Le principe est de construire des patrons d'extraction (extraction patterns) à partir des sorties de l'analyse syntaxique de surface. Il s'agit d'utiliser les informations que l'on a sur les mots pour définir des concepts par des expressions régulières. L'inconvénient de cette approche est que la réalisation des règles d'extraction est manuelle, coûteuse en temps. Elle peut être génératrice de silence si les règles ne couvrent pas tous les cas possibles. Elle demande une grande expertise linguistique.

2.2 Classification

Lorsqu'on cherche à identifier les principales thématiques abordées dans un ensemble de documents, deux approches [4] sont possibles pour regrouper des documents similaires:

- la classification **supervisée** ou catégorisation qui consiste à identifier la classe d'appartenance d'un objet à partir de certains traits descriptifs. Cette approche permet le **classement** automatique de documents dans des classes préexistantes (connues à l'avance), comme par exemple les rubriques d'un journal (société, sport, politique). Les méthodes les plus efficaces sont basées sur un **corpus d'apprentissage [7]**. Elles permettent de déterminer automatiquement la catégorie d'un document à partir d'échantillons de documents représentatifs pour chaque catégorie choisie. …)
- la classification **non supervisée** des documents ou **clustering [6]**, c'est à dire la découverte de classes de documents sans a priori, (on ne connaît pas les classes à l'avance). Comme dans toute approche non supervisée, ces méthodes supposent le choix :
    - d'une représentation des objets à classer (ici les documents sont représentés sous forme de vecteurs[3])
    - d'une mesure de similarité entre les objets
    - d'un algorithme de classification (hiérarchique ou non hiérarchique)

3 Le processus de text mining dans un projet d'intelligence client

3.1 Description du problème à résoudre

Nous décrivons ici le processus mis en place dans le cadre d'un projet d'un constructeur de voitures. L'objectif est de collecter, analyser plus efficacement les retours-clients sur les nouveaux modèles pour les dispatcher le plus rapidement possible aux services concernés, charge à ces derniers de trouver les réponses adéquates aux problèmes soulevés. Lors de la production d'un nouveau modèle, ll est en effet très important de déterminer le plus rapidement possible d'où proviennent les dysfonctionnements ou problèmes mentionnés par les clients (quelles sont les parties de la chaîne de production ou de la chaîne de distribution impliquées dans le problème).
Les retours client sur les nouveaux modèles sont collectés régulièrement via un centre d'appel. Chaque retour est transcrit (verbatim) et stocké dans une base de données. Chaque année, ce sont des dizaines de milliers de verbatim qui sont ainsi collectés.

---

[3] Simple à mettre en œuvre, basé sur une représentation mathématique bien établie, le modèle vectoriel fournit une mesure immédiate de la similarité entre deux vecteurs..

Traditionnellement, ces verbatim sont classés manuellement selon un plan de classification qui doit permettre de faire un contrôle qualité sur toute la chaîne logistique. Plus de trois cents critères ou catégories de problèmes sont ainsi codifiés (cf figure 3.). Cette classification manuelle est très coûteuse en temps (100 verbatim/j/personne), si bien que la base de verbatim initiale ne couvrait qu'un tiers des catégories possibles.

.

AC-01		MAUVAISE PERFORMANCE DE LA CLIMATISATION (CHAUD/FROID)
AC-02		DEFAUT DE FONCTIONNEMENT DE LA CLIMATISATION
AC-03		BRUIT DE FONCTIONNEMENT DE LA CLIMATISATION
AC-05-1		AGREMENT DE MANOEUVRE DES TOUCHES DE CLIMATISATION
AC-05-2		AGREMENT DE MANOEUVRE DES COMMANDES DE CLIMATISATION
AC-06		BRUIT A L UTILISATION DES COMMANDES ET TOUCHES DE CLIMATISATION
AC-07		ODEUR DESAGREABLE DE LA CLIMATISATION
AC-08		IMPRECISION DE L INDICATEUR DE TEMPERATURE EXTERIEURE
AC-09		MAUVAISE VISIBILITE DES TEMOINS LUMINEUX DE LA CLIMATISATION
AC-10		DEFAUT D EVACUATION D EAU DU SYSTEME D AIR CONDITIONNE
AC-11		DEFAUT DE LA FONCTION DESEMBUAGE
AC-12		DEFAUT D AFFICHAGE DE L ECRAN A CRISTAUX LIQUIDES
AC-13		DIFFICULTE DE MANOEUVRE ET BRUIT DES AERATEURS ET VOLETS D ORIENTATION
AC-14		DEFAUT DE FONCTIONNEMENT ET BRUIT DU PURIFICATEUR D AIR
AC-XX		AUTRES DEFAUTS DE LA CLIMATISATION

.

Figure 3 : exemple de code catégorie et sous catégories : AC - Catégorie Climatisation

3.2 Analyse du problème

Comme bien souvent dans les applications industrielles, les données à traiter sont incomplètes ou bruitées [2]. Dans le cas présent, la couverture de la base de verbatim est partielle, De plus, le nombre de verbatim par catégorie est très faible, ce qui rend impossible l'utilisation directe d'une technique de classification supervisée par apprentissage, qui semble a priori le mieux adaptée pour ce type de problème.

Rappelons que le processus pour l'entraînement d'un catégoriseur comporte 3 étapes principales :

- Constituer un corpus d'apprentissage
- Calculer le modèle d'apprentissage pour classer les documents
- Evaluer le modèle d'apprentissage.

Disposer d'un bon corpus d'entraînement est donc essentiel. L'ensemble d'entraînement doit être suffisamment grand pour être représentatif des textes à classer. Une vingtaine de documents par catégorie constitue un ordre de grandeur moyen. Si le nombre d'exemples est trop petit, le catégoriseur risque de retenir des mots spécifiques au corpus d'entraînement et ne sera pas en mesure de procéder à une classification de nouveaux documents.

Le temps des experts étant compté, il est difficile d'obtenir d'eux qu'il fournissent une vingtaine de documents par catégorie. Comment constituer un corpus d'apprentissage en mobilisant le minimum de ressources humaines ?

L'idée a été d'utiliser une technique de clustering non hiérarchique de type KMeans [4, 6] sur la totalité du corpus de verbatim pour grouper les verbatim de contenus similaires. Le schéma ci-dessous illustre les résultat du clustering. Chaque verbatim est représenté par 1 point, dont la couleur indique le code attribué manuellement ; les cercles non colorés sont les documents non classifiés. Les clusters sont symbolisés par un cercle jaune et numérotés de 1 à 9.

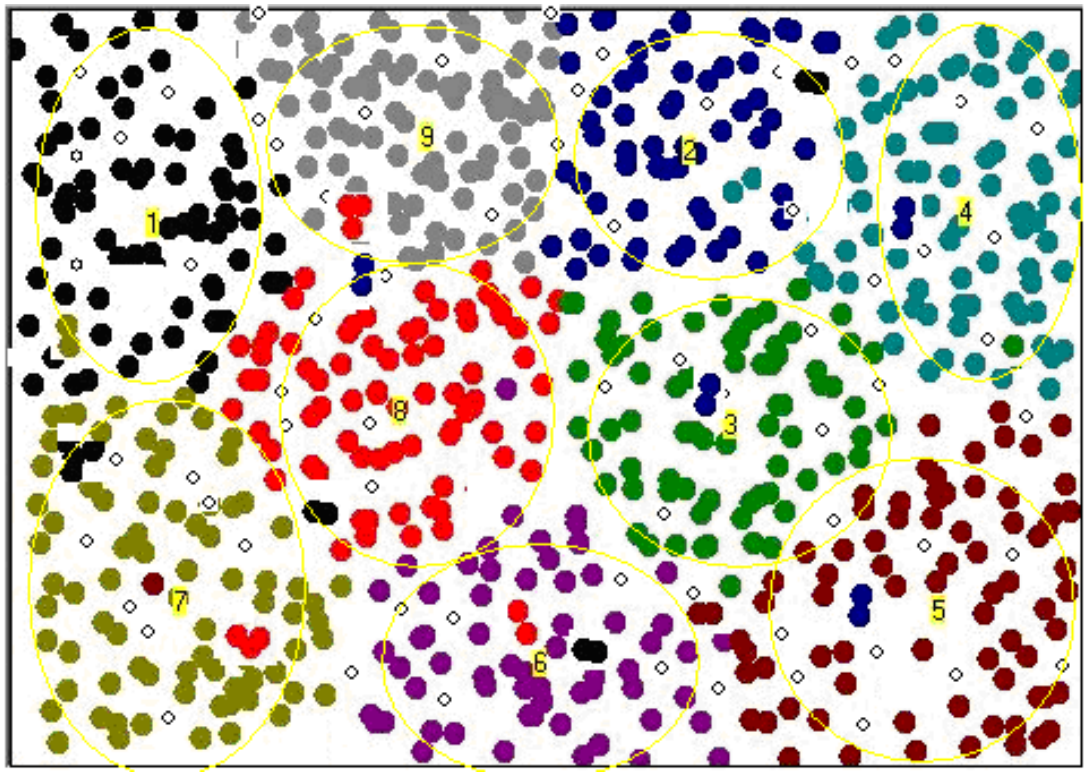


Figure 4 : illustration des résultats du clustering

Dans les clusters résultants, on peut donc trouver aussi bien des verbatim ayant déjà un code de classement attribué manuellement et des verbatim non classifiés. Pour chaque cluster de verbatim, l'expert doit décider si les codes manuels présents pouvait également indexer les verbatim non classifiés. Cette approche a permis de réduire le coût de construction de la base d'apprentissage d'un facteur 3.5

Une fois la base d'apprentissage constituée, il devient alors possible d'employer une technique de catégorisation supervisée de type Rochio [7]. A partir du corpus d'apprentissage, le catégoriseur va chercher à établir des relations entre les termes décrivant les documents et leur catégorie d'appartenance. Il s'agit d'obtenir une sorte de « vecteur prototypique » des catégories ou classes. Ce vecteur est bien sûr construit à l'aide du corpus d'apprentissage. On peut le voir comme un document fictif correspondant à une moyenne des documents de la catégorie.

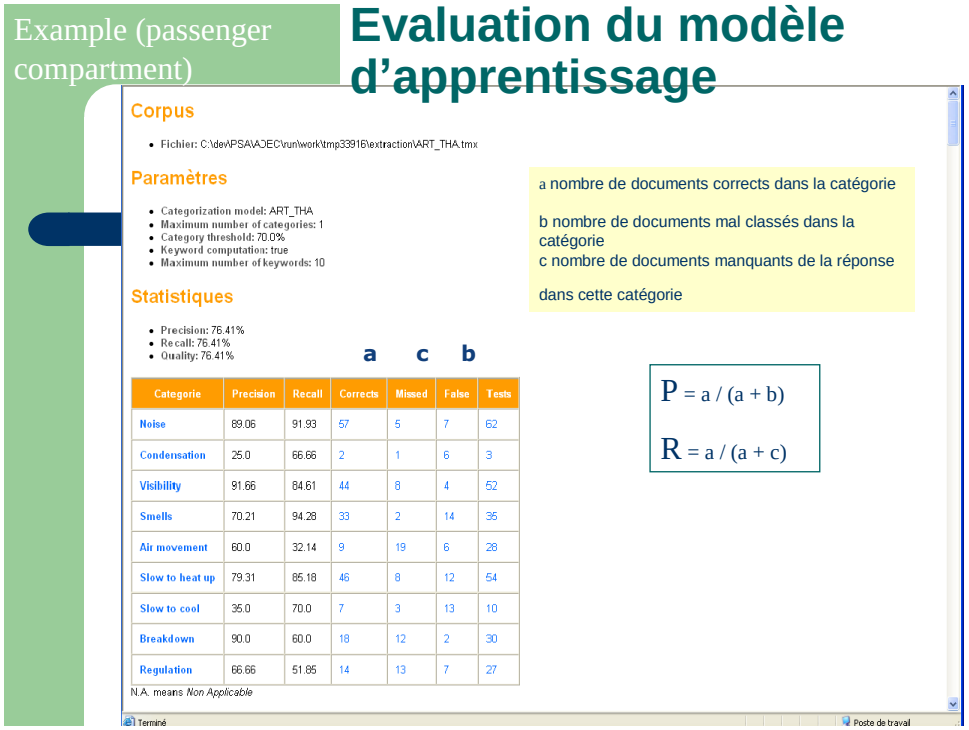
Quand un nouveau document doit être classé, il est alors comparé à ces vecteurs-*type*, en pratique une matrice de poids (mot X catégorie) qui constituera le modèle d'apprentissage. Le classement de nouveaux documents s'opère en calculant la distance euclidienne entre la représentation vectorielle du document et celle de chacune des classes ; le document est assigné à la classe la plus proche.

Souvent masquée par les systèmes de catégorisation automatique, la phase de sélection des termes décrivant les documents est une phase clé car elle fournit le vecteur de termes nécessaire pour la catégorisation. Pour l'extraction des termes, l'emploi d'un analyseur morpho-syntaxique[4] d'est révélé bénéfique car il permet d'éliminer certaines catégories de mots (les prépositions, les articles, …), et de s'affranchir des problèmes de variations morphologiques des noms, adjectifs, verbes, etc.
Pour la sélection elle-même, le principe général a été d'utiliser une méthode statistique (calcul de l'information mutuelle, méthode du chi-2) pour sélectionner les termes qui ont une corrélation avec les catégories en se limitant à ceux qui ont une fréquence suffisante pour éviter d'avoir un modèle trop spécifique du corpus d'apprentissage. Généralement, on élimine également les mots peu fréquents.

En aval de la catégorisation par apprentissage, il a été parfois nécessaire de définir des règles d'exploitation de la catégorisation afin d'éliminer certaines doubles catégorisations ou au contraires autoriser la catégorisation multiple sur des catégories qui se chevauchent naturellement dans le domaine d'application concerné. L'ajout de ressources linguistiques propres au domaine traité (dictionnaires, thésaurus) peut améliorer la catégorisation permettant ainsi au système de faire la différence entre un problème dû à une coulure de peinture (incident qui peut survenir dans la chaîne de production) ou à une griffure (incident qui peut survenir dans la chaîne de distribution).

L'ensemble du processus est itératif. Au fur et à mesure de l'utilisation du catégoriseur sur de nouveaux documents, il faut évaluer périodiquement le modèle à partir d'une sélection de nouveaux documents classés pour éviter le risque d'une dégradation des performances. Cette évaluation s'effectue classiquement par des mesures de rappel et de précision [4], comme le montre la figure 5 ci-dessous.

[4] Les analyseurs morpho-syntaxiques sont de plus en plus performants (20 Mo/h pour Xelda de Xerox Research Centre Europe) et capables de traiter de plus en plus de langues différentes. Citons également Nooj, une plate-forme linguistique, freeware issue d'INTEX (dévellopé par Max Silbertzein).



Categorie	Precision	Recall	Corrects	Missed	False	Tests
Noise	89.06	91.93	57	5	7	62
Condensation	25.0	66.66	2	1	6	3
Visibility	91.66	84.61	44	8	4	52
Smells	70.21	94.28	33	2	14	35
Air movement	60.0	32.14	9	19	6	28
Slow to heat up	79.31	85.18	46	8	12	54
Slow to cool	35.0	70.0	7	3	13	10
Breakdown	90.0	60.0	18	12	2	30
Regulation	66.66	51.85	14	13	7	27

Figure 5 : évaluation du modèle d'apprentissage

3.3 Bénéfices

Outre le gain de temps humain, l'automatisation du processus de classification des verbatim permet une vérification en continu de l'adéquation des catégories par rapport au ressenti client, par exemple :

- Identification des items ne faisant pas l'objet de plainte et proposition de suppression d'items
- Identifications de plaintes non classées pouvant entraîner la création de nouveaux items

Une fois classés, les verbatims peuvent faire l'objet d'une analyse plus fine, par exemple du lien entre la fréquence / gravité des plaintes et d'autres facteurs (code du problème, kilométrage du véhicule au moment du problème, …).

4 Conclusion

Nous avons montré sur un exemple comment des outils de text mining pouvaient permettre d'analyser plus efficacement des retours-clients. D'une manière générale, un système intégrant des composants de text mining peut permettre :

- d'extraire des concepts similaires dans un corpus multi-lingue (rassemblant des documents en différentes langues),
- découvrir les sujets traités dans un corpus (clustering),
- classer des documents selon une taxonomie, ...
- calculer des distributions d'entités ou de concepts extraits (noms de personnes, compagnies, lieux, produits, …)

Qu'il s'agisse de veille, de gestion de la relation-client ou de gestion de connaissances, le rôle d'outils ou de systèmes de text mining [12] est essentiellement de supprimer les tâches les plus fastidieuses (extraire des informations, trier, classer) pour libérer du temps pour des opérations où la créativité intellectuelle est essentielle (analyse des besoins, analyse systémique, mise en contexte de l'information, synthèse, mise en perspective, …).
Le tableau ci-dessous donne quelques exemples de besoins ou de problèmes où l'emploi de techniques de text mining permet d'apporter des solutions. Ils sont classés selon le type de traitement à effectuer. [5]

CRM Customer Relationship Management : gestion de la relation client
♣Comprendre les préférences/demandes des clients, en analysant des questions ouvertes ou les notes prises au niveau d'un centre d'appel, en vue d'anticiper un départ chez le concurrent, une action en justice, proposer une nouvelle offre, …
♣+♦Orienter les mails clients reçus sur le site vers les services adéquats et les aider à répondre le plus rapidement et correctement possible

IE : Intelligence économique, veille économique, technologique et environnementale

[5] Un logo indique le type de traitement à effectuer :

♣	Extraction d'information
♦	Catégorisation ou classification supervisée
♥	Clustering ou classification non supervisée
♠	Bibliométrie

♣identifier les actions ou faits relatifs aux stratégies des entreprises: les noms des acteurs, compagnies, organisations, personnes, associés à des informations financières (chiffres d'affaires, rentabilité, croissance), commerciales (parts de marché, nombre de clients, nouveau client), boursières (capitalisation, tendances), des stratégies d'entreprises (prise de participation, fusion, acquisition, ouverture de filiales, création de joint venture), etc.
♣+♦Classer des news, rapports, brevets, etc. selon des rubriques métiers ou selon des profils de veilleurs
♣+♥Anticiper des risques (émeutes, crises financières, catastrophes, …) en analysant la presse locale d'un pays
♣+♥identifier les actions ou faits relatifs aux personnes ou organisations citées dans des rapports de police, établir des corrélations entre des lieux, des types d'évènements, des armes, des drogues, …
♣+♥+♠Identifier les thématiques de recherche sur un domaine, identifier les relations existants entre les acteurs (co-auteurs, citations, co-citations) de la recherche à partir de la littérature scientifique et technique telle qu'elle est représentée dans les bases de données
♣+♥+♠protéger son image et sa réputation en analysant les propos tenus, les opinions émises sur les médias et notamment Internet et les forums
♣Analyser la portée juridique d'un brevet en extrayant automatiquement les revendications indépendantes, les liens avec d'autres brevets,
…

KM : gestion des connaissances
♣+♥Identifier les experts de l'entreprise et leurs champs d'expertise à partir des documents qu'ils produisent et diffusent sur l'intranet
♣+♦Classer automatiquement des documents selon les référentiels métiers de l'entreprise
♣Analyser des CV pour en extraire des compétences dans des secteurs d'activités
♣Identifier les modifications ou les renvois à un texte juridique …

Pour autant, il est rare que les données à traiter permettent une utilisation directe des outils. Il faut le plus souvent, pour transformer l'information brute en information utile, exploitable, imaginer des processus de traitement adaptés aux réalités du terrain (données, bruitées, manquantes, temps disponible, …). C'est sans doute cela le quotidien d'une ingénierie des connaissances.

Bibliographie

1. Grishman, R. *Information Extraction: Techniques and Challenges*. In Pazienza, M. T., editor, Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, Frascati, Italie. LNAI Tutorial, Springer, 1997.

2. Grivel Luc " Customer Feedbacks and Opinion Surveys Analysis in the Automotive Industry ", chapter 13 in [12]

3. Grivel Luc. « Outils de classification et de catégorisation pour la fouille de textes » , ISKO – Semaine de la connaissance, SDC 2006, ISKO - Pratiques et méthodes de classification du savoir à l'heure d'Internet - 26 juin **2006**.p. 95-104

4. Grivel Luc "Intégration de Composants de Text Mining pour le développement d'un système de recherche et d'analyse d'information" , **Les systèmes d'information élaborée**, Ile Rousse, Corse, Edition CD-ROM **2002** (CRRM - Marseille)**,,** article accessible sur www.isdm.org *ISDM Information Sciences for Decision Making* 2003. N°6

5. Lewis D. *Feature Selection and Feature Extraction for Text Categorization*. Proceedings of Speech and Natural Language Workshop, pp. 212-217, 1992.

6. Mandreoli F Martoglia R & Tiberio T. *Text clustering as a mining task*, chapter 3 in [12]

7. Rocchio J.-J. *Relevance Feedback in Information Retrieval*, p. 313–323. in The SMART Retrieval System : Experiments in Automatic Document Processing, G. Salton (editor), Prentice-Hall Inc. : New Jersey, 1971.

8. Salton G., McGill M. *Introduction to Modern Information Retrieval*. McGraw-Hill. 1983

9. Sebastiani F. *Text categorisation*, chapter 4 in [12]

10. Wilks Y., *Information Extraction as a Core Language Technology*. In Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, ed. M. T. Pazienza, Frascati, Italy, LNAI Tutorial, Springer-Verlag. p. 14-18, 1997.

11. Yang Y. Peterson J. *Feature Selection in statistical learning of text categorisation*. Proceedings of the fourteenth international conference on Machine Learning, pp 412-420, Nashville, 1997.

12. Zanasi A. (Editor) Text Mining and its Applications To Intelligence, Crm And Knowledge Management, ISBN 1-85312-995-X Series: Management Information Systems Vol 9, 327p. WIT Press, 2005.