Analyse en ligne de l'information : une approche permettant l'extraction d'informations stratégiques basée sur la construction de composants de connaissance

Luc Grivel, Sylvie Guillemin-Lanne, Pascal Coupet, Charles Huot

Analyse en ligne de l'information : une approche permettant l'extraction d'informations stratégiques basée sur la construction de composants de connaissance

Grivel Luc, Guillemin-Lanne Sylvie, Coupet Pascal, Huot Charles

{luc.grivel, sylvie.guillemin-lanne, pascal.coupet, charles.huot} @temis-group.com

TEMIS **Text Mining Solutions** 59, rue de Ponthieu 75 008 Paris

Mots clefs :
Fouille de données textuelles, extraction information, traitement du langage naturel, classification, catégorisation, intelligence économique
Keywords:
Text mining, information extraction, natural language processing, clustering, categorization, business intelligence
Palabras clave :
Explotación del texto, extracción de la información, proceso natural de lenguaje, clasificación, inteligencia de negocio

Résumé

Les entreprises réclament des systèmes d'extraction et d'analyse de l'information **personnalisés** et **évolutifs** mettant l'accent sur des besoins relatifs à des fonctions (commercial, recherche, technique) et des profils industriels (automobile, banque, etc.) : identifier des fournisseurs de technologies, des concurrents, repérer les fusions/acquisitions d'entreprises paraissant dans la presse économique, etc. Pour extraire les aspects essentiels des documents et les mettre en relation avec la demande des utilisateurs, il est indispensable de combiner diverses techniques provenant de domaines comme l'analyse de données, la linguistique, l'intelligence artificielle. Chacune de ces techniques est vue comme un composant aux fonctionnalités précises et délimitées.

Pour illustrer le développement de tels systèmes d'aide à l'analyse, l'article détaille le fonctionnement de ces composants au sein d'un logiciel de veille, Online Miner™ qui intègre :

- des composants de connaissance liés au domaine d'application et aux objectifs de veille.
- Un administrateur de sources d'information
- un serveur d'extraction
- un serveur de recherche documentaire
- un serveur incorporant un moteur de classification et un moteur de catégorisation
- des composants statistiques graphiques permettant de construire tableaux et cartes.

Les avantages de cette approche sont à la fois techniques et fonctionnels. Plus simples à développer, plus robustes parce que testés dans des contextes différents, ces composants de text mining peuvent s'assembler de plusieurs manières pour créer de manière économique et fiable des systèmes d'aide à l'analyse **personnalisés et évolutifs.** Au niveau fonctionnel, les apports essentiels sont la prise en compte de l'utilisateur (son domaine d'application, la diversité de ses objectifs d'analyse (veille, CRM, KM, …), l'automatisation de l'alimentation du système, les modes d'accès et d'analyse de l'information sont pris en charge par des composants spécialisés.

1 Introduction

Que cela soit au niveau d'un individu, d'une entreprise ou d'une nation, anticiper les évolutions de son environnement est vital pour maintenir ou développer sa compétitivité. L'information est au cœur d'une telle démarche d'Intelligence Economique. En effet, un grand nombre de documents publiques et disponibles sur Internet (dépêches de presse, bases de données bibliographiques scientifiques et techniques, …) ou en Intranet (mails électroniques, rapports techniques, rapports d'étonnement) contiennent potentiellement de l'information utile à la décision. 80% de cette information en ligne est textuelle.

Notre postulat est qu'aucun outil, aucune technique, aucun algorithme ne peut résoudre seul les problèmes liés à :

- l'hétérogénéité des informations, que cela soit d'un point de vue :
    - contenu sémantique : financier, commercial, technique, …
    - structurel : fortement structuré (brevet) à non structuré (e-mails)
    - linguistique (multilinguisme)
    - format du support (Word, html, pdf, …)
    - taille : définition de l'unité d'information à analyser (granularité de l'information)
- le volume croissant d'information.
- la diversité d'expression des besoins et des points de vue des consommateurs d'information. Chacun aimerait pouvoir accéder à l'information qui l'intéresse dans les documents, avec les concepts de son métier, les problématiques de consommation d'information relative à sa fonction, selon son propre schéma de lecture, un peu comme lorsque nous soulignons dans un texte les mots ou passages qui nous importent.

Pour extraire les aspects essentiels des documents et les mettre en relation avec la demande des utilisateurs, il est indispensable de combiner diverses techniques provenant de domaines comme l'analyse de données, la linguistique, l'intelligence artificielle. L'ensemble de ces techniques appliquées au texte constitue ce qu'on appelle le text mining, cœur du métier de la société TEMIS. Les techniques employées s'appuient sur l'expérience acquise antérieurement par les membres de l'équipe qui avaient développé 'Intelligent Miner For Text' et 'Technology Watch' pour IBM ou « Sdoc » et « Henoch » pour l'INIST-CNRS [COUPET 1995, 1998], [HUOT 1992,1998], [GRIVEL 2000, 2001].

L'objectif de cet article est de décrire une architecture permettant de combiner les techniques d'analyse du texte (segmentation, lexicale, syntaxique, sémantique) et diverses techniques d'accès à l'information (index, classification, catégorisation, cartographie). Chacune de ces techniques est vue comme un composant aux fonctionnalités précises et délimitées. Plus simples à développer, plus robustes parce que testés dans des contextes différents, ces composants peuvent s'assembler de plusieurs manières pour créer ainsi des applications variées et adaptées.

Cette architecture est doublement validée :
1) par **une application industrielle** que nous avons **menée avec Telcal/Intersiel et IBM Italie.** Il s'agit d'un système de surveillance de la concurrence dans les domaines de l'agriculture, de l'artisanat et du tourisme. Cf **Alessandro Zanasi (IBM Market Intelligence Italie)** dans l'article intitulé **Text Mining : The New Competitive Intelligence Frontier. Real Application Cases in Industrial, Banking and Telecom/SMEs World**
2) par un logiciel, **Online Miner™** développé par TEMIS qui intègre et « met en scène » les composants de text mining. Ce logiciel sera présenté en section 2.

2 Online Miner : application de collecte et d'analyse en ligne de documents textuels

2.1 Architecture

L'architecture repose sur trois principes de conception qui sont autant de manières permettant d'assembler nos composants de text mining pour atteindre l'objectif fixé (développer des applications permettant d'extraire les aspects essentiels des documents et les mettre en relation avec la demande des utilisateurs) :

- **architecture client/serveur** : chaque composant constitue un serveur basé sur la technologie RMI (Remote Method Invocation) et dispose d'une API (Application Programmatic Interface) Java publique lui permettant de s'intégrer facilement dans une application existante.
- **modélisation sémantique** des documents par des meta-données provenant de l'application de différentes techniques d'analyse. Le fait de disposer d'un modèle sémantique garantit de pouvoir utiliser un langage de requête évolué de type SQL.
- **XML,** comme langage pivot de description de documents échangés entre les différents serveurs.

Dans le cadre de cette architecture, développer une chaîne de traitement de l'information consiste à définir des composants « données » (fichiers de configuration décrivant les caractéristiques de chaque type de source (Web, fichier XML, répertoire, …) et leurs modes d'accès), les composants « méthodes d'analyse » à utiliser, les composants « métiers » relatifs au domaine d'application (appelés ici composants de connaissance ou Skill Cartridges™), les composants d'interface utilisateur pour construire tableaux et cartes, et composants de pilotage pour décrire l'enchaînement des tâches à effectuer.

Dans ce processus itératif qu'est la veille, on peut distinguer quatre fonctions essentielles d'un système d'analyse de l'information :

- Automatiser la constitution et la mise à jour régulière d'une base documentaire, avec la meilleure couverture possible pour les axes de surveillance recensés. De cette manière sont constituées des bases de références pour les différents acteurs de l'Intelligence Economique et Stratégique de l 'entreprise.
- Annoter les documents par extraction d'information en vue de leur accès et leur organisation ultérieure.
- Les stocker dans une base documentaire
- Fournir une interface conviviale permettant d'exploiter cette base selon différents modes de recherche et scénarios d'analyse, en combinant recherche, statistiques, catégorisation et classification du résultat de la recherche.
-

Comment ces fonctions sont-elles prises en charge ?

La figure 1 montre comment, en intégrant uniquement les composants nécessaires, il est possible d'étendre un système de recherche d'information existant par une automatisation de la collecte en amont, et, en aval, par des mécanismes de classification ou de catégorisation de l'information.

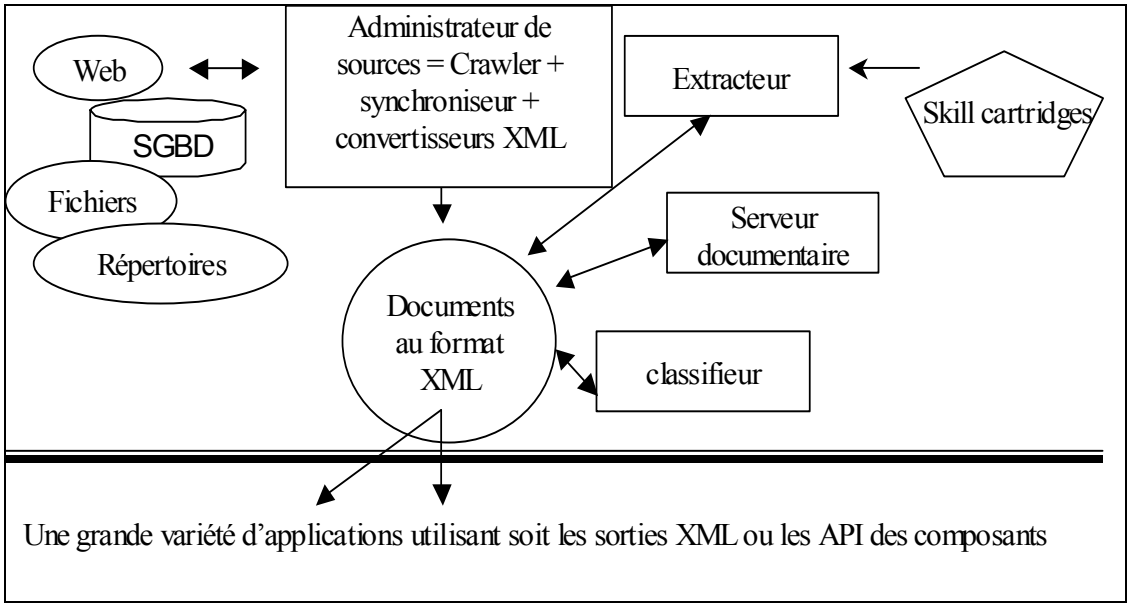

FIGURE 1

1) L'administrateur de sources 'crawle' différentes sources d'information, actionne des filtres de conversion XML des documents (en fonction du type de document et de la date de mise à jour) et les stocke dans un répertoire local.
2) Un serveur d'extraction basé sur la technologie des transducteurs[1] utilise des composants de connaissance ('Skill Cartridge™ ') pour dégager les concepts clé contenus dans les documents (noms de compagnies, dates, valeurs monétaires, fonctions, lieux, ou tout autre concept relatif à un domaine…) et génère les metadonnées décrivant chaque document.
3) Un serveur de recherche documentaire stocke et indexe les documents et leurs metadonnées.
4) Un serveur incorporant un moteur de classification et un moteur de catégorisation classifie les documents (constitue des groupes), ou les catégorise (les place dans des groupes définis à priori).

Pour détailler ces composants et montrer comment ils interagissent dans cette architecture, nous les présentons à travers une application basée sur ce principe : Online Miner™.

2.2 L'administrateur de sources

Il met à jour une base de données à partir de documents provenant de différentes sources pré-définies. Ce composant inclut les fonctions permettant de se connecter à différentes sources d'information, (Web, mail, news, data banks) dans différent formats (ascii, HTML, Word, PowerPoint, Excel, PDF) et de définir un profil de recherche sur ces sources. Tous les documents correspondant à ce profil sont périodiquement et automatiquement recherchés, copiés et stockés dans un répertoire local. Ils sont ensuite convertis dans un format homogène XML en veillant à qualifier de manière unique les éléments et les attributs. En effet, à partir du moment où de plus en plus de données collectées auprès de serveurs sur le Web s'expriment en XML, les risques de collision de noms deviennent plus importants. C'est pourquoi nous utilisons la notion d'espace de noms d'XML (http://www.mutu-xml.org/xml-base/shared/KEY-NAMESPACES.html) et les éléments du Dublin Core pour décrire les meta-données (**http://purl.org/dc/elements/1.1/**).

[1] Le langage de description des règles d'extraction est de la famille des langages réguliers. Les transducteurs (automates d'états finis) sont un moyen efficace d'opérer sur ces langages. Cette efficacité garantit de pouvoir traiter de gros volumes de données.

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<doc_list xmlns="http://temis-group.com/docRef/"
        xmlns:dc="http://purl.org/dc/elements/1.1/">
<DOC>
     <dc:Identifier>6a62355891d2f65343fbcd912bf2243e</dc:Ident
     ifier>
  <dc:Date>2001/08/01</dc:Date>
  <dc:Title>Russia - Birthplace of Supercomputer.</dc:Title>
  <dc:Source>NOVECON 01/08/2001</dc:Source>
  <dc:Creator />
  <Country>RUSSIA</Country>
  <domains />
  <text>Russian scientists created and launched a supercomputer
     operating at a speed of 1 teraflop (1,000 billion operations
     with a floating point per second). The new machine is
     designated MBS-1000, is based on 768 Alpha processors with
     667 GHz frequency, has 768 Gbyte RAM and a 7,680 Gbyte
     disk memory. The superproductive computer was created by
     the KVANT Research Institute and other institutes.
     Development was coordinated by the interdepartmental
     computer center with the financial support of the Industry
     and Science Ministry (R230 million), the Russian
     Technological Development Fund (R87m), the Russian
     Fundamental Research Fund (R27m) and the Russian
     Academy of Sciences (R50m). Another R20m was received
     from the federal program titled State Support for the
     Integration of Higher Education and Fundamental Science.
     Russian natural monopolies displayed interest in the
     supercomputer. Talks on cooperation are conducted by
     BOEING and ENERGIA. A contract to sell the computer to
     India may also be signed. Researchers are working on a 5
     teraflop computer, MBS 5000, scheduled for commissioning in
     2003. The new computer based on 1,500 processors will cost
     another $20m. Source: VREMYA NOVOSTEI, August 1,
     2001.</text>
  <Copyright>(c) Novecon 2001.</Copyright>
</DOC>
</doc_list>
```

2.3 L'extracteur et les composants de connaissance (Skill Cartridge™)

Outre la récupération des meta-données existantes (par exemple, les champs qui sont présents dans le cas de documents provenant d'une base bibliographique), il est indispensable de générer des meta-données décrivant le contenu du document pour implanter un mode de recherche. C'est ici qu'intervient la phase d'extraction d'information, qui peut aller, selon la Skill Cartridge™ employée d'une simple extraction d'éléments syntaxiques comme les noms, les verbes, à une extraction d'éléments sémantiques tels que des noms de compagnies, des relations (fusion de X avec Y, achat de W par Z), des noms de lieux, des dates, des prix…

Une *Skill Cartridge*™ est une hiérarchie de composants de connaissance. Un composant de connaissance peut avoir la forme d'un dictionnaire de termes ou d'un ensemble de règles d'extraction. Une règle d'extraction s'exprime sous forme d'une expression régulière pouvant comporter des

lemmes (forme canonique des mots), des étiquettes syntaxiques et sémantiques. Le mécanisme est décrit dans [Gri01].

Le serveur d'extraction est capable de traiter 7 langues (anglais, français, allemand, espagnol, hollandais, portugais, italien).

Sur le plan technique, le serveur d'extraction procède par étapes :

- segmentation des données textuelles
- lemmatisation et
- tagging (à chaque mot est affecté une catégorie syntaxique)
- application des Skill Cartridges™
  - comparaison des lemmes du texte aux termes décrits dans les dictionnaires
  - recherche des patterns décrits dans les règles et affectation aux éléments trouvés d'une étiquette sémantique.

Le résultat de l'extraction est une annotation du document par les concepts extraits (FIGURE 2).

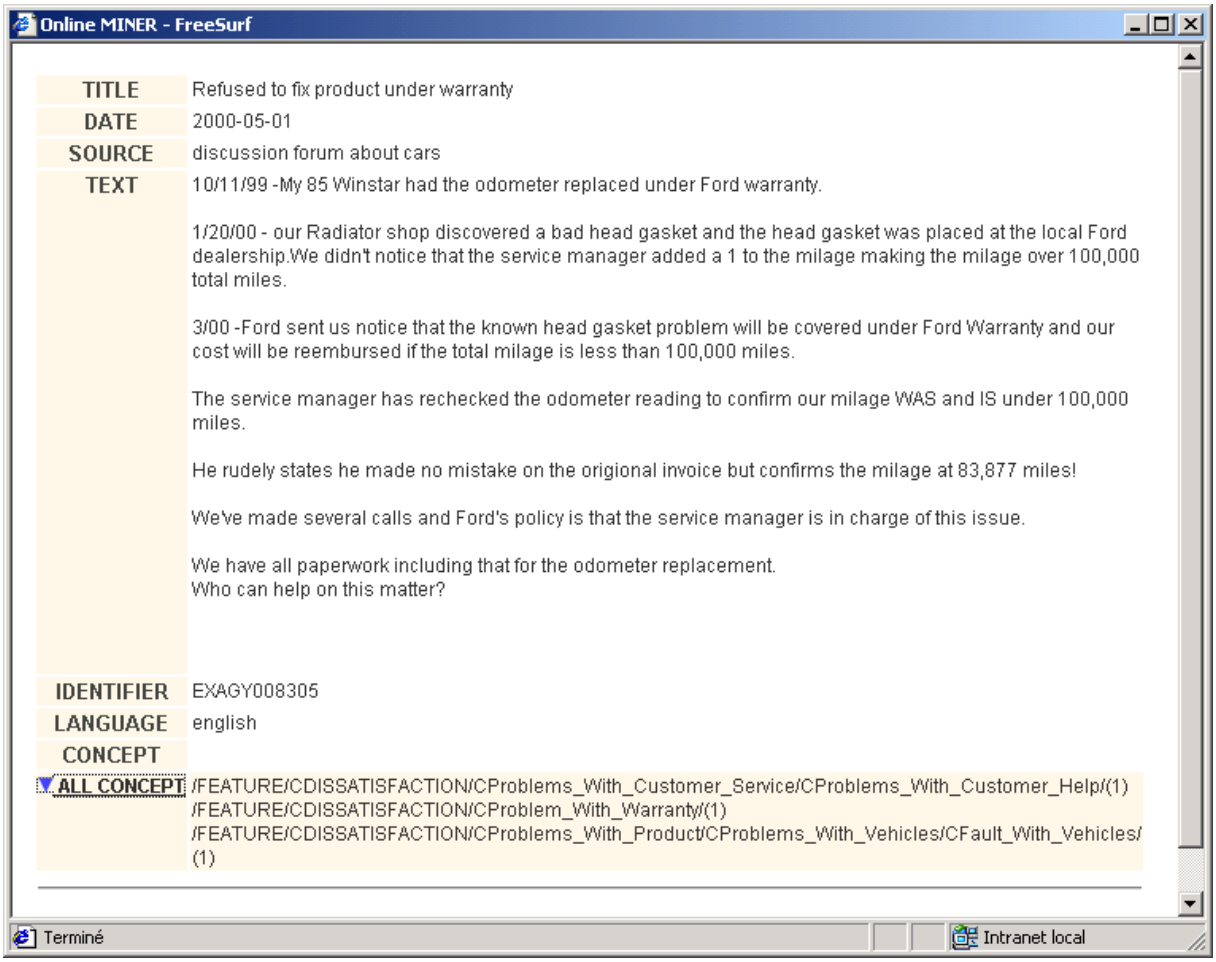

FIGURE 2

ou, suivant l'utilisation ultérieure, une liste de « slots » exprimés sous forme d'attribut valeurs (FIGURES 3 et 4).

```
Annonce d'une intention (rumeur) de prise de participation d'une compagnie
dans une autre:

France Telecom said it would acquire a 54.3% stake in Equant
```

```
/CI_Extraction
      France Telecom say it would acquire a 54.3 stake in equant
      France Telecom said it would acquire a 54.3% stake in Equant
   who:    France Telecom     France Telecom
            /Telecom_Operator    France Telecom     France Telecom
   which_information:   say it would acquire a 54.3 stake
                        said it would acquire a 54.3% stake
      Announcement:     say     said
      Rumor:            would   would
      /CI_Financial/buying_stake    acquire a 54.3 stake
                                    acquire a 54.3% stake
         /buying_acquisition        acquire          acquire
         /CI_Financial/stake        a 54.3 stake     a 54.3% stake
            how_much_percent:       a 54.3           a 54.3%
   whom:    equant     Equant
            /Telecom_Operator    equant    Equant
```

FIGURE 3

Analyse de contenu de e-mails:

Mon téléphone à reçu un choc et depuis l'alimentation est souvent de mauvaise qualité ou se coupe.
Mon capital de points pour le renouvellement est de + 7000 points or le premier téléphone que vous proposez est à près de 7800 points.

J'aimerais que vous me fassiez une proposition de renouvellement rapide quitte à payer un léger surcoût afin de pouvoir changer mon téléphone le plus rapidement possible.

Je dispose actuellement d'un NOKIA dont je suis satisfait et j'aimerais un NoKIA de la nouvelle gamme.

```
~DISSATISFACTION/~Problems_With_Objects/~Faulty_Objects
      téléphone à reçu un choc      téléphone à reçu un choc
   WHAT_OBJECT:   téléphone         téléphone

~DISSATISFACTION/~Problems_With_Objects/~Faulty_Objects
      mauvais qualité ou se couper        mauvaise qualité ou se coupe

~NOTIFICATIONS    point pour le renouvellement  points pour le
renouvellement
   WHAT_OBJECT:   point       points
   WHAT_ACTION:   ~Buy  renouvellement    renouvellement

~NOTIFICATIONS    aimer que vous me faire un proposition de renouvellement
rapide                 quitte_à payer un léger surcoût
                  aimerais que vous me fassiez une proposition de
renouvellement                rapide quitte_à payer un léger surcoût
   WHAT_ACTION:   ~Buy  renouvellement    renouvellement

~NOTIFICATIONS    changer mon téléphone   changer mon téléphone
   WHAT_ACTION:   ~Buy  changer           changer
   WHAT_OBJECT:   téléphone         téléphone

~SATISFACTION/~Satisfied_with_Products
                  NOKIA dont je être satisfaire       NOKIA dont je suis
satisfait
   WHAT_OBJECT:   NOKIA             NOKIA
```

```
~NOTIFICATIONS/~Notifications_Other
                  je aimer un NoKIA de la nouvelle gamme
                  j' aimerais un NoKIA de la nouvelle gamme
   WHAT_OBJECT:   NoKIA             NoKIA
```

FIGURE 4

2.4 les composants 'méthodes d'organisation des données' et représentation graphique

A l'issue de l'étape d'extraction, chaque document est donc annoté par des concepts. Les informations de catalogage (auteur, source d'information, date, …) quand elles existent, constituent des caractéristiques utiles pour l'organisation des données et leur accès par les utilisateurs.

Plusieurs approches sont possibles et peuvent être combinées selon l'objectif recherché (faciliter l'analyse et la synthèse de l'information ou faciliter la recherche):

1. Recherche par mots-clés ou sur des concepts exprimés dans la/les Skill Cartridges™
2. Classification: la classification permet de regrouper les documents similaires par thèmes sans a priori sur la structure thématique.
3. Cartographie,
4. Catégorisation de documents : sur la base d'un vecteur de caractéristiques comprenant, entre autres, les résultats de l'extraction, la catégorisation permet de router les documents vers des rubriques ou catégories prédéfinies.
5. Analyse statistique

Trois exemples sont présentés.

1er exemple : router des documents vers des rubriques ou catégories prédéfinies

La Catégorisation ou classification supervisée peut être un des moyens pour prendre en compte un profil d'utilisateur. Ce dernier exprime en effet de manière explicite ses centres d'intérêt en fournissant un corpus de documents dont il a validé lui-même la catégorisation. Une vingtaine de documents de taille moyenne par catégorie suffit généralement.

L'utilisation du composant de catégorisation peut se faire en amont du stockage sur le serveur documentaire des documents et leurs meta-données ou de manière dynamique sur le résultat d'une recherche.

Le calcul d'un modèle d'apprentissage (qui sera le profil généré pour l'utilisateur) est réalisé en prenant diverses mesures de l'apport d'une caractéristique à l'ensemble des catégories en fonction de différents paramètres :

- Le nombre de catégories pour un document (affectation à 1 ou plusieurs catégories)
- Le nombre minimal de documents dans une catégorie
- La fréquence des caractéristiques dans le document
- La fréquence des caractéristiques dans une catégorie
- …

L'apprentissage s'effectue en deux étapes qui consistent à :

- sélectionner les caractéristiques pertinentes pour les catégories en utilisant les échantillons,
- attribuer à chaque caractéristique un poids qui sera utilisé pour l'affectation des documents supplémentaires.

Cette matrice des poids (caractéristique X catégorie) constitue le modèle d'apprentissage.

L'évaluation de la qualité du modèle construit se fait sur un ensemble de documents du corpus annoté qui n'ont pas participé à l'apprentissage (par défaut 10% du corpus annoté de départ) : mesure de la précision et du rappel à partir du nombre de documents corrects dans la catégorie, du nombre de documents manquants de la réponse dans cette catégorie, du nombre de documents mal classés dans la catégorie et du nombre de documents du corpus de test appartenant à la catégorie.

Une fois l'apprentissage réalisé, pour affecter de nouveaux documents aux catégories, il suffit de sélectionner un modèle d'apprentissage et le fichier de documents supplémentaires à catégoriser. Un seuil de catégorisation permet de ne garder dans une catégorie que les documents dont la valeur dans cette catégorie est supérieur à ce seuil.

2ème exemple : combiner la classification et la cartographie

Une telle combinaison permet de réduire le temps passé par l'utilisateur pour trouver les documents qui l'intéressent lorsque la réponse à une requête comporte trop de documents pour être analysée séquentiellement. Grâce à cette fonction, le résultat d'une requête peut être éclaté en un petit nombre (fixé d'avance, usuellement entre 5 et 10) de groupes de documents sémantiquement proches qui sont ensuite projetés sur une carte thématique. L'utilisateur peut sélectionner le thème qui lui semble le plus pertinent et demander à nouveau son éclatement en sous-groupes si nécessaire.

L'avantage de cet approche est que l'utilisateur n'a pas à exprimer des requêtes complexes, nécessitant de connaître toutes les finesses du langage de requête et des plans de classement éventuellement utilisés. Il peut exprimer des requêtes plus simples et s'appuyer sur la classification pour sélectionner les documents les plus intéressants : il classifie les documents à la demande (FIGURE 5), navigue de thèmes en thèmes à travers une carte globale, visualise les relations inter-thèmes sur cette carte (FIGURE 6), accède à la liste de documents du thème, etc.

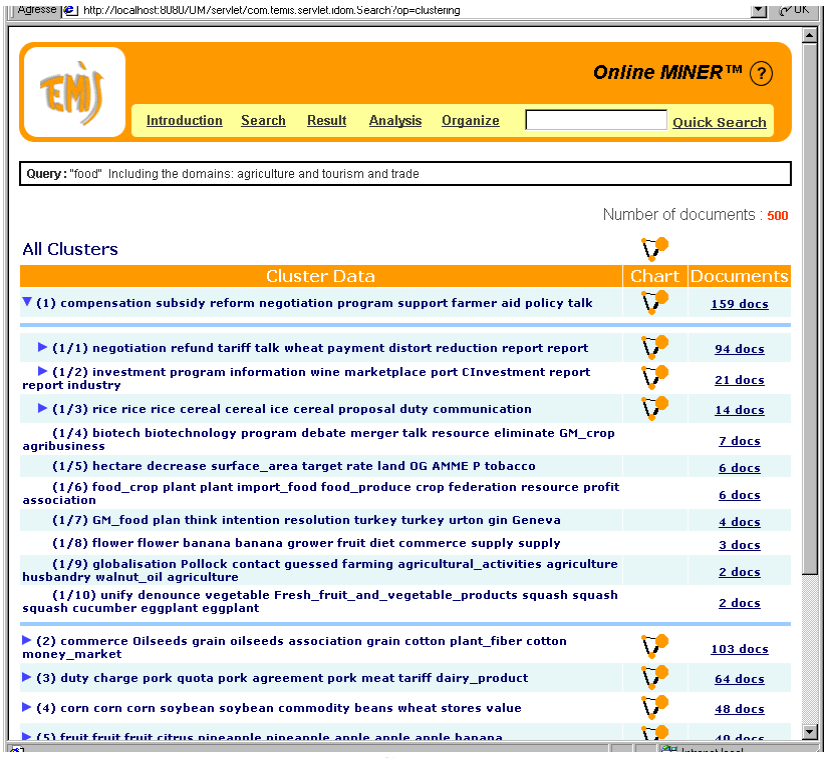

FIGURE 5

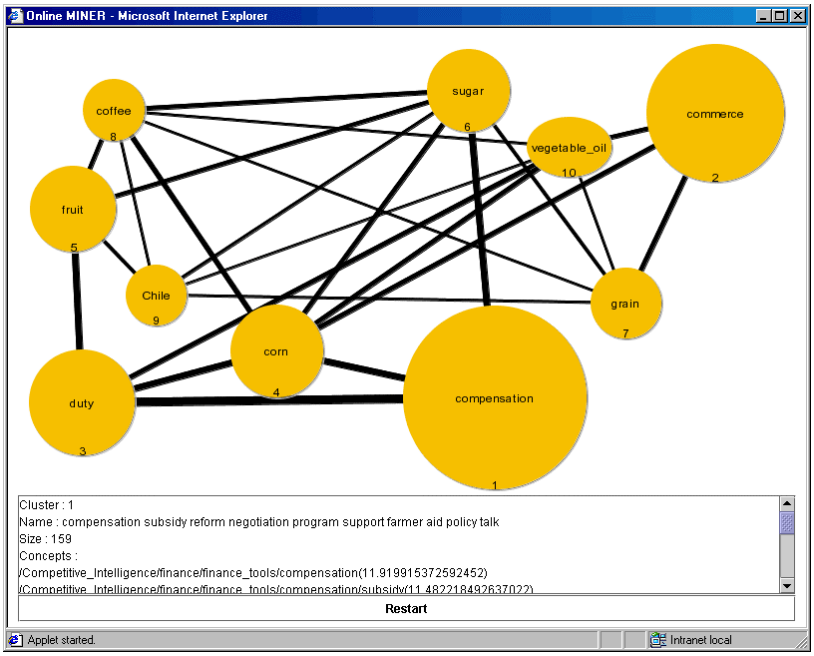

FIGURE 6

3ème exemple. Analyse statistique

L'analyse statistique a pour objectif de fournir une vue globale d'un ensemble de documents (résultat d'une requête, résultat d'une classification, …) pour évaluer sa pertinence. Une vue globale de la distribution des documents en fonctions des dates, relations identifiées ou concepts est fournie sous forme graphique ou tabulaire. Les deux exemples ci-dessous (FIGURES 7 et 8) sont des illustrations d'applications de respectivement une Skill Cartridge™ « Intelligence économique » et d'une Skill Cartridge™ « Analyse de la satisfaction »

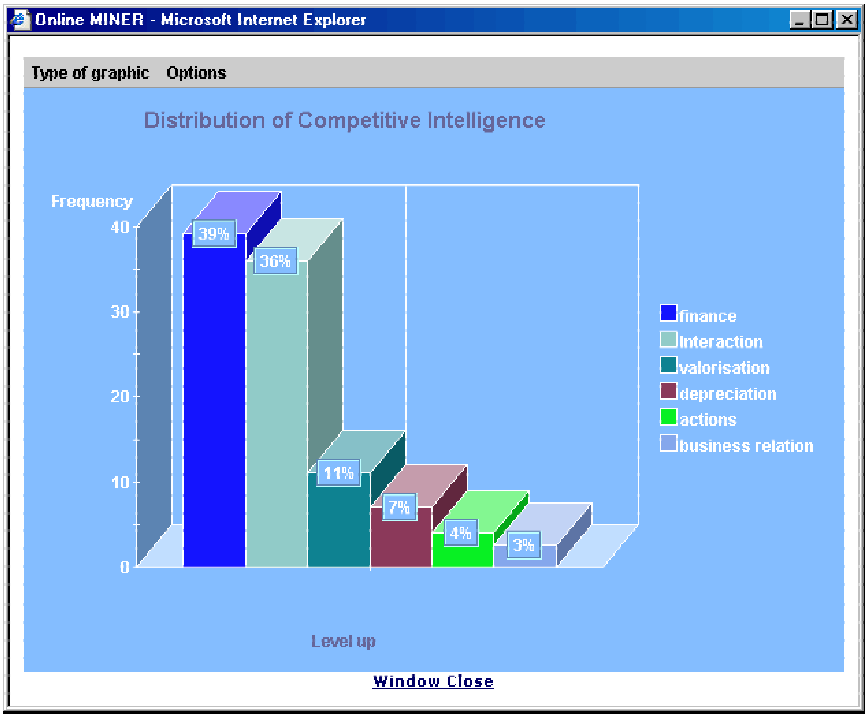

FIGURE 7

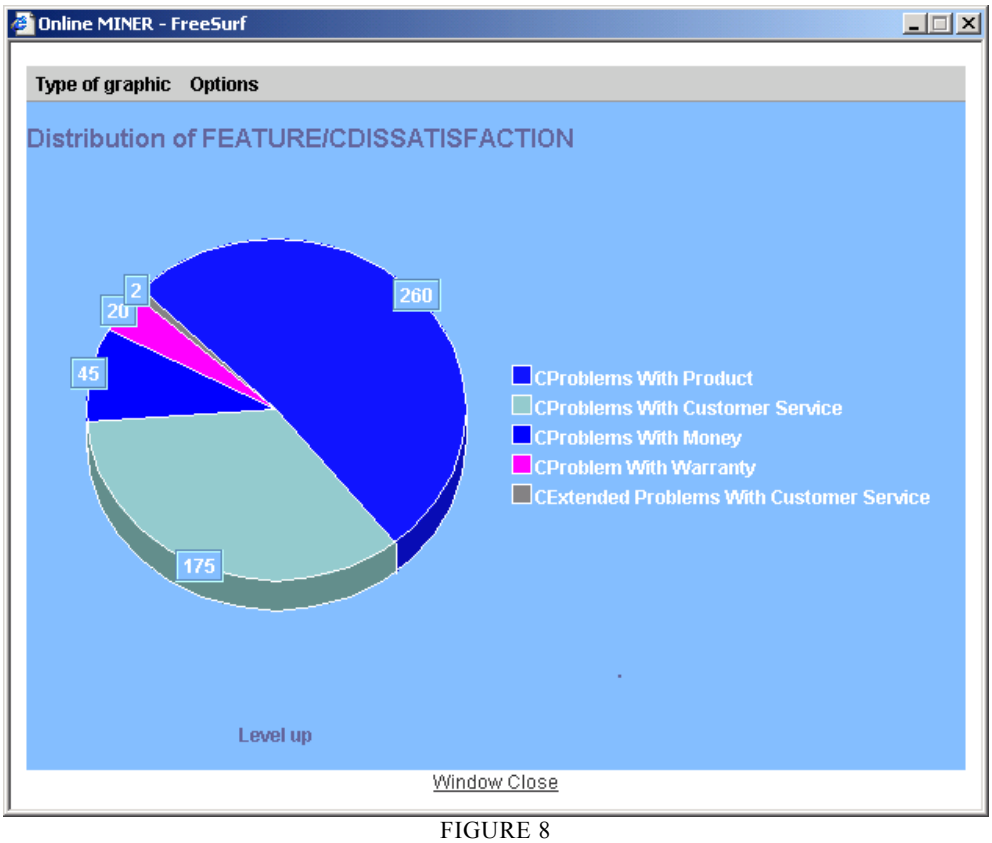

FIGURE 8

3 Conclusion

Notre stratégie est de développer des composants de text mining qui puissent facilement s'intégrer dans toute application visant à extraire les aspects essentiels d'un texte ou d'une collection de textes. Les avantages de cette approche sont multiples. Plus simples à développer, plus robustes parce que testés dans des contextes différents, nos composants de text mining peuvent être assemblés pour créer des applications variées et adaptées. Une telle architecture permet de développer de manière économique et fiable des systèmes d'aide à l'analyse **personnalisés et évolutifs.** Pour paraphraser François Bourdoncle dans une interview récente, ce type d'architecture permet un passage à l'échelle pour pouvoir construire des systèmes fonctionnant sur de gros flux de documents : chaque serveur spécialisé, développé dans un langage moderne, (RMI java, C++) peut fonctionner sur des machines différentes.

Au niveau fonctionnel, outre l'automatisation de la collecte d'information, la prise en compte du profil de l'utilisateur et de la diversité des objectifs d'analyse (veille, CRM, KM, …) en fonction de la nature de données à analyser sont certainement les points les plus importants. Comme nous l'avons souligné dans la section 3.2, notre approche met l'accent sur l'utilisation des composants de connaissance pour prendre en compte le domaine d'application, l'objectif d'analyse et la nature des données. Un point clé est la maintenance des Skill Cartridges™ par l'entreprise utilisatrice. Un environnement de maintenance des composants de connaissance relatifs aux besoins stratégiques particuliers de l'entreprise, qui, par nature, sont confidentiels et non diffusables, est en cours de développement. Décrit sur le plan fonctionnel dans [Grivel et al. 2001], il offrira une interface de mise à jour **des Skill Cartridges™** i.e. la terminologie et les règles d'extraction d'information relatives à un axe d'analyse, et permettra le **partage des Skill Cartridges™ entre les ingénieurs qui les développent**.

Si une telle architecture garantit des capacités d'intégration rapides, il faut encore que les composants soient de qualité et correspondent à l'état de l'art. Prenons comme exemple **le multi-linguisme**, souvent indispensable. Ici un fort partenariat avec le **Centre de Recherche Xerox** de Grenoble nous a permis de répondre à ce besoin. TEMIS consacre environ 12% de son budget à la recherche, accueille actuellement une stagiaire de DEA du **CRRM Centre de Recherche Rétrospective de Marseille, Université d'Aix Marseille III** et, last but not least, s'est vu décerner récemment le label technologie-clé par l'**ANVAR**.

Bibliographie

[ASH 92] R. Ashori-Hechmati, JP. Chanod, S. Guillemin-Lanne: "Analyse syntaxique globale et correction grammaticale étendue", *Actes de la 12ème conference internationale AVIGON'92, Artificial Intelligence, Expert systems, Natural Language*, conférence spécialisée sur le traitement du langage naturel et ses applications (volume 3), Avignon, juin 1992.

[BOU 99]Bourigault D. et Jacquemin, C. (1999) : Term Extraction + Term Clustering: An Integrated Platform for Computer-Aided Terminology. *In Proceedings, 8th Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, Bergen, pages 15-22.

[BOU 98] Bourigault D. & Habert B. (1998). Evaluation of Terminology Extractors: Principles and Experiments, In Proceedings of the First International Conference on Language Resources and Evaluation, vol. I, pp. 299-305, A. Rubio, N. Gallardo, R. Castro and A. Tejada, Editors, Granada, Spain.

[BOU 00] Bourigault D. & Slodzian M. (2000) *Pour une terminologie textuelle*, Terminologies Nouvelles, n° 19

[CHA 92] J.-P. Chanod, M. El-Beze, and S. Guillemin-Lanne. Coupling an Automatic Dictation System with a Grammar Checker. In *Proc. of the 14th COLING*, pp. 940-944, Nantes, France, 1992.

[CHA 00] Charlet J., Zaklad M., Kassel G. et Bourigault D. (eds.) *Ingénierie des Connaissances, évolutions récentes et nouveaux défis*. Paris : Eyrolles, 2000.

[CON 00] Condamines A.et Aussenac-Gilles N. : Entre textes et ontologies formelles : les bases de connaissances terminologiques. I*n Capitalisation des connaissances*. Zacklad M .Grundstein M.(Eds.).Paris : Hermès. Traités IC2000.

[COU 95]Coupet P., Grandjean N., Huot C.et Chellali T.: Application du logiciel Technology Watch à l'analyse du développement pharmaceutique pour le domaine des maladies neurodégénératives, Les systèmes d'information élaborée, *Congrès S.F.B.A.,* juin 1995.

[COU 98]Coupet P., Hehenberger Michael.: Text Mining applied to patent analysis, Les systèmes d'information élaborée, *Annual Meeting of American Intellectual Property Law Association (AIPLA) Airlington.,* octobre 1998.

[FEL 97] Fellbaum, C.: A Semantic network of English Verbs. In C. Fellbaum (ed.) *WordNet: an Electronic Lexical Database*. Cambridge MA: MIT Press (1997).

[FEL 99] Fellbaum, C. : La représentation des verbes dans le réseau sémantique WordNet. *Langages 136*. (1999).

[FAU 00] FaureD. Poilbeau T. Extraction d'information utilisant INTEX et des connaissances sémantiques apprises par ASIUM, premières expérimentations, RFIA 2000, F. Schmitt et I Bloch Editeurs, Paris, France.

[GRI 97]Grivel L., Polanco X., Kaplan A. : 'A computer System for Big Scientometrics at the Age of the World Wide Web', *Scientometrics*, vol.40, n°3, 493-506, 1997.

[GRI 95] Grivel L., Mutschke P., Polanco X.: 'Thematic mapping on bibliographic databases by cluster analysis : a description of SDOC environment with SOLIS', *Journal of Knowledge Organization*, Vol. 22, n°2, 70-77, 1995.

[GRI 95] Grivel L., Francois C. : 'Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et technique' - *Solaris* n°2 "Les sciences de l'Information : Bibliométrie, Scientométrie, Infométrie", Presses universitaires de Rennes, p.81-113, 1995.

[GRI 00] Grivel L. : L'hypertexte comme mode d'exploitation des résultats d'outils et méthodes d'analyse de l'information scientifique et technique, thèse de doctorat en Sciences de l'information et de la communication, Université Aix-Marseille III,. 10 janvier 2000.

[GRI 01] Grivel Luc, Guillemin-Lanne Sylvie, Lautier Christian, Mari Alda La construction de composants de connaissance pour l'extraction et le filtrage de l'information sur les réseaux Filtrage et résumé automatique de l'information sur les réseaux, 3ème congrès du Chapitre français de l'ISKO International Society for knowledge Organization, 5-6 juillet 2001

[HUO 92] Huot C., Quoniam L., Dou H.: New method concerning analysis of downloaded data for strategic decision, *Scientometrics*, Vol 4, n° 2, pp 279-294, 1992

[HUO 98] Huot C.: Text mining solutions, *The Journal of Association for Global Strategic Information*, Vol 7, issue 1, march 1998

[JAC 94] Jacquemin, C.. FASTR : A unification grammar and a parser for terminology extraction from large corpora. In *Proceedings, Journées IA'94*, pages 155-164, Paris. Paris : EC2. (1994e)

[KAR 00] Karttunen Lauri Applications of Finite-State Transducers in Natural Language Processing *In: Proceedings of CIAA-2000. Lecture Notes in Computer Science. Springer Verlag.*.

[MAN 99] Manning C.D. et Schütze H. : *Foundations of Statistical Natural Language Processing*. Cambridge MA: MIT Press. (1999*)*.

[MAR 97] Mari, A. et Saint-Dizier, P. (1997). Générativité : au delà d'une théorie des types. Grenoble, *TALN'97*.

[MCC 00]McCallum, A.K. et al. (2000). Automating the Construction of Internet Portals with machine Learning. *Proc. COLING'00*.

[NAZ 97] Nazarenko A., Zweigenbaum P., Bouaud J., Habert B., Corpus-Based Identification and Refinement of Semantic Classes, *Journal of the American Medical Informatics Association*, vol. 4 (suppl), 585-589. 1997.

[OGO 94]Ogonowski et al. (1994) Tools for Extracting and Structuring Knowledge from Texts. *Proc COLING-94*.

[RIL 99]Riloff E. Jones R. Learning Dictionnaries for Information Extraction by multi level Bootstrapping Proceedings of Sixteenth National Conference on Artificial Intelligence, AAAI 1999, Orlando Floride.

[SOD 99] Soderland S. Learning Information Extraction Rules for Semi-Structured and Free Text, Machine Learning 34, 233-272, 1999.

[TOU 98] Toussaint, Y., Namer, F., Daille, B., Jacquemin, C., Royauté, J. et Hathout, N. (1998). *Une approche linguistique et statistique pour l'analyse de l'information en corpus*. In P. Zweigenbaum, editor, *Proceedings, TALN'98*, pages 182-191.