La construction de composants de connaissance pour l'extraction et le filtrage de l'information sur les réseaux

une approche industrielle

Grivel Luc, Guillemin-Lanne Sylvie, Lautier Christian, Mari Alda

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

{luc.grivel, alda.mari, sylvie.guillemin-lanne, christian.lautier}
@temis-group.com

Résumé :

L'accroissement de l'activité économique, scientifique jointe à l'éclosion des nouvelles technologies de l'information se traduisent par une croissance remarquable de l'information disponible sous forme électronique et exigent de nouveaux outils capables d'analyser et de structurer des documents textuels afin de permettre à des utilisateurs non-experts de fouiller ces documents et/ou de les évaluer.

Notre article présente les premiers résultats d'un programme de recherche et de développement de la société TEMIS, dont le cœur d'activité est le **text mining**, c'est-à-dire l'extraction, le traitement, la visualisation et la capitalisation de l'information à partir des données, produites ou reçues, par une société dans son domaine.

L'objet de ce programme est le développement d'une station de travail appelée Knowledge Station ou K-Station, indépendante du domaine d'application et multilingue. La K-Station automatise un processus itératif de construction et validation de composants de connaissances, i.e. la terminologie et les règles d'extraction d'informations relatives à un secteur d'activité, et ce, à partir de corpus de données textuelles. Elle a pour objectif de permettre à un linguiste de construire un ensemble de *composants de connaissances* pour un secteur industriel déterminé (automobile, chimie…) ou un métier particulier (journalisme, informatique, …).

Nous mettrons l'accent sur le processus d'extraction de l'information à partir de corpus, en insistant sur la méthodologie utilisée. Notre approche est fondée sur la réutilisation de bases terminologiques existantes et sur l'acquisition de connaissances à partir d'un corpus selon un processus itératif intégrant des règles d'extraction. Elle repose sur une intégration d'outils dans une architecture client/serveur. Nous présentons les différentes étapes d'analyse intégrées dans cette architecture. Notre exposé sera illustré par des exemples de résultats dans le domaine de l'intelligence économique. Les *composants de connaissance* s'intègrent automatiquement dans une suite logicielle nommée *Insight Discoverer*™ .

Mots-clés : fouille de données textuelles, ingénierie des connaissances, extraction d'information, règle d'extraction, base de connaissances

Abstract :

With the steady growth of business and scientific activities and the recent advances in Information Technology, huge amounts of electronically available but unstructured data have to be dealt with. New tools able to analyse and structure textual data need to be developed so that non-expert users can understand and evaluate the contents of their documents.

This paper describes the first results of a R&D program set up by TEMIS Company, whose core business is Text Mining, i.e. information extraction, information processing, visualisation and valorisation of all the data issued by or received in a company in its field of activity.

The aim of the program is to build a multilingual knowledge station, called K-Station, in order to build knowledge components (terms and extraction rules) in an iterative way. The knowledge station is independent of the field of activity.

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

The K-Station allows the linguist to create knowledge components for a determined industrial sector (cars or chemistry for example) or for a specific job (e.g. journalism or computer science).

We particularly insist on the process of information extraction from a corpus, by detailing the methodology used. Our approach is based on the recycling of existing terms databases and knowledge discovery from a corpus, according to an iterative process which includes extraction rules. All the components are part of a client/server architecture. We explain the various levels stages of analysis integrated within this architecture and we illustrate them with results taken from the field of Competitive Intelligence. Knowledge components are included in a software suite called *Insight Discoverer*™.

Keywords : Text Mining, knowledge ingeniering, information extraction, extraction rule, knowledge discovery

1. Contexte

Cet article présente les premiers résultats d'un programme de recherche et de développement d'une société européenne récemment créée, TEMIS. Le cœur d'activité de cette société est le développement de logiciels de text mining, c'est-à-dire de logiciels permettant l'extraction, le traitement, la visualisation et la capitalisation de l'information à partir des données, produites ou reçues, par une société dans son domaine.

L'environnement de text mining mis en place implémente 4 fonctions principales :

1. Collecte et sélection de corpus de données textuelles
2. Extraction d'information à partir de ce corpus
3. Organisation des documents du corpus par thèmes, en mode supervisé (catégorisation) ou non supervisé (classification)
4. Visualisation et synthèse des résultats

Parmi ces quatre fonctions, la deuxième requiert la création et la validation de ressources terminologiques et de règles d'extraction spécifiques au domaine d'application. C'est la phase la plus coûteuse du processus. Son automatisation est donc essentielle. Comment la réaliser ? Dans un contexte industriel, l'objectif est d'obtenir un modèle d'extraction (des règles et des termes), qui s'intègre parfaitement à l'application du client (ses sources d'information usuelles, ses données).

Dans cet article, nous décrivons le processus d'extraction de l'information mis en place. Nous en déduisons quelques éléments pour assister la constitution de règles d'extraction et définir les premières bases d'un programme de recherche et développement.

L'objet de ce programme est le développement d'une station de travail à usage interne appelée Knowledge Station ou K-Station. La K-Station permettra à un linguiste de construire un ensemble de *composants de connaissances* pour un secteur industriel déterminé (automobile, chimie…) ou un métier particulier (journalisme, informatique, …). Actuellement, le linguiste n'est pas guidé/assisté dans l'enchaînement des tâches à réaliser. La K-Station automatise un processus itératif de construction et validation de composants de connaissances, i.e. la terminologie et les règles d'extraction d'informations relatives à un secteur d'activité, et ce, à partir de corpus de données textuelles.

Cette recherche s'appuie sur l'expérience acquise antérieurement par les membres de l'équipe de R&D, qui en tant que responsables d'activités de text mining pour une grande société d'informatique ou pour l'INIST-CNRS, ont mené à bien plus de 200 applications de text mining dans des secteurs d'activité divers (finance, chimie, administration publique, commerce, biologie, physique, etc.) [BEDECARRAX 1993], [COUPET 1995], [GRIVEL, 1995,1997, 1999, 2000], [HUOT 1992,1998].

2. Méthode d'extraction d'information

Notre approche est fondée sur l'utilisation de bases terminologiques existantes et sur l'acquisition de connaissances à partir d'un corpus selon un processus itératif. Elle s'inscrit dans le courant BCT, bases de connaissances terminologiques (groupe TIA, Groupe Ingénierie des Connaissances GRACQ)[ABEILLE 1997], [AUSSENAC, 2000], [BOURIGAULT 1998, 1999, 2000], [CHARLET 2000], [CONDAMINES 2000], [HABERT 1998], [NAZARENKO 1997], [JACQUEMIN 1994]. Elle exploite les complémentarités de différents processus d'analyse du texte : analyse morpho-syntaxique, extraction (de noms propres, de relations, …) à partir de marqueurs linguistiques.

L'idée de base est de construire des patterns[1]. Par exemple, si l'objectif est de détecter dans des dépêches de presse, les fusions ou acquisitions entre des sociétés, on recherchera des expressions comme '*company* acquired *company' or 'company* merged with *company'*.

company peut être le nom d'une société mais aussi tous les mots qui peuvent faire référence à une société (firm, manufacturer, …), ce qui implique de les regrouper sous un descripteur sémantique '*company*'. Pour une application d'intelligence économique , la classe des verbes indiquant un transfert de possession contiendra des verbes comme buy, hold, acquire, …

Dans une dépêche de presse, les acteurs économiques sont mentionnés dans un contexte particulier. Il est nécessaire d'associer à la classe des mots regroupés sous un même descripteur les spécifieurs la qualifiant. On le fera à l'aide 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/ou sémantiques. Dans l'exemple cité, un acteur comme *the first private label pasta maker* pourra être capturé par la règle :

```
(company_Adj|Loc_Adj|#ORD)* / (brand_product) / (Company)[2]
```

Enfin une expression comme '*company* acquired *company'* pourra être détectée par une règle de type :

```
company / possession_transfer / company
```

A ce niveau, il est alors possible de définir les rôles (who et which_company) des acteurs concernés en tenant compte de l'information syntaxique (voie active vs passive, modalités, auxiliaires et négations) :

```
{company:who} / (HAVE|MODAL)* / possession_transfer / {company :which_company}[3]
```

Cette dernière règle permet de dégager des rôles remplis par who et which_company (qui achète qui).

```
company_acquisition     Spigadoro , Inc. Acquires Largest European Private
Label Pasta Company

   who            Spigadoro , Inc.  [883,897]
   which_company  Largest European Private Label Pasta Company
```

Le serveur d'Extraction procède en deux étapes que nous allons détailler : l'analyse morpho syntaxique - à chaque mot est affectée une catégorie syntaxique - et l'application des règles d'extraction.

---

[1] Un pattern est une expression régulière qui identifie le contexte de syntagmes pertinents et les délimiteurs de ces syntagmes. En associant une action à un pattern, une règle ajoute de l'information à une séquence de mot, par exemple en lui attribuant un nom de classe sémantique qui peut être ensuite utilisé dans d'autres règles.

[2]
Company_Adj renvoie aux adjectifs tels que leading, industrial, public
LocAdj → des adjectifs de lieux tels que European, Italian
#ORD → ordinaux (first, second)
brand_product → private label, branded + #NOUN

[3] Cette règle détectera: Company1 would acquire Company 2, Company1 acquired Company 2, Company1 has bought Company 2, mais non: Company2 has been acquired by Company 1.

1. *Analyse morpho-syntaxique*. Elle affiche en sortie l'étiquette morpho-syntaxique de chaque mot et est couplée avec un modèle statistique (chaîne de Markov), lorsqu'il y a ambiguïté sur les parties du discours.

```
A leading manufacturer of branded products in the Mediterranean food sector, today
announced that it will acquire Pastificio Gazzola S.p.A., leading manufacturer of
private label pasta.

A                       <AT>                    <a>
leading                 <NN>                    <leading>
manufacturer            <NN>                    <manufacturer>
of                      <IN>                    <of>
branded                 <VBN>                   <brand>
products                <NNS>                   <product>
in                      <IN>                    <in>
the                     <AT>                    <the>
Mediterranean           <JJ>                    <Mediterranean>
food                    <NN>                    <food>
sector                  <NN>                    <sector>
,                       <CM>                    <,>
today                   <NR>                    <today>
announced               <VBD>                   <announce>
that                    <CS>                    <that>
it                      <PPS>                   <it>
will                    <MD>                    <will>
acquire                 <VB>                    <acquire>
Pastificio              <NP>                    <Pastificio><guessed>
Gazzola                 <NP>                    <Gazzola><guessed>
S.p.A.                  <NP>                    <S.p.A><guessed>
,                       <CM>                    <,>
leading                 <NN>                    <leading>
manufacturer            <NN>                    <manufacturer>
of                      <IN>                    <of>
private                 <JJ>                    <private>
label                   <NN>                    <label>
pasta                   <NN>                    <pasta>
```

Figure 1

2. *Analyse sémantique*. Des classes sémantiques sont formées grâce à :
   a. Des dictionnaires, en exploitant des ressources existantes comme WordNet [FELDBAUM 1997, 1999] pour l'Anglais, et le thesaurus développé à l'université de Caen par l'équipe de Victorri, disponible à l'adresse http://elsap1.unicaen.fr , pour le Français.

What dictionary	*Action dictionary*
<brandname_> <Mineral_water>	<communication_> <announcement_>
Arrowhead Badoit Blaue / Quellen Buxton Calistoga Contrex Vittel (Source)? / Perrier Poland / Spring Quézac Salvetat ...	announce announcement notify notification proclamation promulgation statement talk ...

Figure 2

b. un ensemble de règles contextuelles, combinant lemmes, syntaxe et sémantique. Ces règles permettent notamment de contrôler localement des phénomènes de polarité, négation, construction de syntagmes ou de circonstants tels que les compléments temporels, extrêmement utiles dans le domaine de l'intelligence stratégique.

```
<actor_in_agribusiness>

        ;; a seed and flour miller
(company_Adj|Loc_Adj|#OD)* / (NOUN)* / (food_|brand_product) / (actor|company)

       ;; maker of private label pasta
(company_Adj|Loc_Adj|#OD)* / (actor|company) / of / (food_|brand_product)
```

Figure 3

Etiquetage sémantique versus étiquetage morpho-syntaxique

```
A leading manufacturer of branded products in the Mediterranean food sector, today
announced that it will acquire Pastificio Gazzola S.p.A., leading manufacturer of
private label pasta.

A               <AT>    <a>
leading         <NN>    <leading>                   <company_Adj>
manufacturer    <NN>    <manufacturer>              <company>
of              <IN>    <of>
branded         <VBN>   <brand>
     products   <NNS>   <product>                   <brand_product>
in              <IN>    <in>
the             <AT>    <the>
Mediterranean   <JJ>    <Mediterranean>             <loc_Adj>
food            <NN>    <food>                      <food_>
sector          <NN>    <sector>
,               <CM>    <,>
today           <NR>    <today>
announced       <VBD>   <announce>                  <announcement_>
that            <CS>    <that>
it              <PPS>   <it>
will            <MD>    <will>
acquire         <VB>    <acquire>                   <possession_transfer>
Pastificio      <NP>    <Pastificio><guessed>
     Gazzola    <NP>    <Gazzola><guessed>
     S.p.A.     <NP>    <S.p.A><guessed>            <company>
,               <CM>    <,>
leading         <NN>    <leading>                   <company_Adj>
manufacturer    <NN>    <manufacturer>              <company>
of              <IN>    <of>
private         <JJ>    <private>
     label      <NN>    <label>                     <brand_product>
pasta           <NN>    <pasta>                     <food_>
```

Figure 4

Le module d'extraction exploitant étiquetage morpho-syntaxique, reconnaissance de classes sémantiques et règles synctatico-sémantiques utilise la technologie des transducteurs [HOBBS 1997], [KARTTUNEN 2000], optimisant ainsi la rapidité du traitement.

A l'issue du traitement, on obtient un texte taggué, exploitable par d'autres applications.

```
actor_                      leading manufacturer of branded products
agribusiness_market         Mediterranean food sector
   food_                    Mediterranean food

communication_/information_/announcement_
                            announced
   when time_/punctual_     today
   what_announcement        company_acquisition

company_acquisition         will acquire Pastificio Gazzola S.p.A.
   who
   which_company            Pastificio Gazzola S.p.A.

actor_in_agribusiness       leading manufacturer of private label pasta
   food_                     private label pasta
```

Figure 5

Quelques remarques avant de passer à la description fonctionnelle de la station de travail que nous voulons développer.

Ecrire des règles demande de réfléchir aux interactions entre des informations de nature différente :

- Des annotations syntaxiques portant sur des mots, des syntagmes nominaux reflétant la structure de parties de phrases
- Des annotations sémantiques reflétant le sens des mots

La difficulté dans cette tâche est de spécifier des règles qui capturent le plus d'expressions correctes possibles (rappel) sans pour autant augmenter le nombre de cas non pertinents (précision).

On peut distinguer deux types d'approches opposées dans l'écriture de patterns :
- partir des expressions pertinentes trouvées dans le corpus d'entraînement et les traduire par des patterns. Cette approche évite la surgénération mais les règles sont souvent trop spécifiques pour s'adapter à un nouveau corpus.
- partir de règles syntaxiques, définir des mini-grammaires locales avant même de regarder le corpus. Cette approche est souvent sur-productive mais plus générale.

Une des idées que nous allons chercher à développer dans la K-Station est de faciliter le passage de l'écriture de règles spécifiques à l'écriture de règles plus génériques en minimisant autant que possible le nombre de cas non-pertinents.

Soulignons que, les patterns n'étant pas tous de même niveaux (du plus général au plus spécifique), ils ne s'appliquent pas tous au même moment . Nous préconisons d'appliquer :

1. les patterns de reconnaissance des noms propres, des valeurs numériques, monétaires, expressions temporelles,
2. ensuite, les patterns qui s'appuient sur l'analyse morpho-syntaxique et qui sont indépendants du domaine,
3. puis les patterns qui sont spécifiques au domaine économique ou scientifique,
4. et enfin, ceux relatifs à l'application visée.

Il faut donc que la K-station permette de gérer des hiérarchies de patterns et de contrôler leur exécution sur des corpus. C'est pourquoi nous définissons la notion de cartouche de connaissance ou *Skill Cartridge*™.

Une *Skill Cartridge*™ est une hiérarchie de composants de connaissance. Un composant de connaissance peut avoir la forme d'un dictionnaire ou d'un ensemble de règles d'extraction. Chaque composant participe à l'extraction.

Sur la base de ces remarques les principales fonctions de la K-station sont :

- **la définition et la création d'un corpus** documentaire à partir de différentes sources d'information, notamment en provenance de l'Internet, via un aspirateur de site (web crawler) capable de récupérer et de mettre à jour les documents souhaités.
- **la construction des composants**. de connaissance ou *Skill Cartridge*™ dont le contenu répond aux questions suivantes:

  - who : Qui sont les acteurs du secteur d'activité ou du domaine économique étudié ?
  - what : Quels sont les objets considérés relatifs au domaine décrit ?
  - how : Quelles sont les actions de ces acteurs, sur quels objets portent-elles et comment s'effectuent-elles ?
  - where : Où ont lieu les actions en question ?
  - when : Quant ont-elles eu lieu ?
  - how much : Et pour quel montant ?

  Les étapes de construction sont détaillées ci-dessous :

  - l'analyse morpho-syntaxique des textes constituant le corpus d'entraînement puis la définition du vocabulaire pertinent relatif au secteur d'activité ou au domaine économique étudié (agroalimentaire, automobile, .etc) figure 1 et 2,
  - le regroupement des termes ainsi définis sous des descripteurs sémantiques eux-mêmes organisés, selon les besoins, en une hiérarchie simple (3/4 niveaux) (figure 3),
  - la définition de règles d'extraction d'information s'appliquant aux concepts de premier niveau (un terme renvoie à un descripteur) (figure 4) afin de dégager des concepts de niveau supérieur et d'aboutir aux concepts visés (figure 5) : les acteurs du domaine d'activité étudié – fabricants, fournisseurs, négociants, multinationales, etc., - leurs actions - rapprochements inter-entreprises, lancement de produits – et toute information relative au lieu, au temps ou correspondant à une donnée financière,
  - l'exécution interactive des règles d'extraction sur le corpus d'entraînement afin d'évaluer le résultat de l'extraction. Ainsi, de manière interactive, il est possible d'enrichir le vocabulaire, de modifier des règles d'extraction et d'en apprécier l'impact sur le corpus de travail. La K-Station présente alors au linguiste les concepts extraits, ce qui lui permet progressivement de **valider les composants** pour un domaine d'application.

Enfin la K-Station permet au linguiste de **partager ses composants** avec d'autres concepteurs de *skill cartridge*™ . Les *skill cartridge*™ sont stockées au format XML selon une DTD décrivant les classes sémantiques, les relations existant entre ces classes et les règles d'extraction.

La K-Station est indépendante du domaine et supporte 7 langues dont 4 sont actuellement exploitées (français, allemand, anglais, italien).

3. Travaux futurs et Conclusion

Pour l'instant, nous nous sommes surtout attachés à définir un environnement de gestion des *Skill Cartridge*™ adapté à notre processus d'extraction. Les temps d'exécution des règles sont encore un peu longs pour permettre une réelle interactivité. Récemment, plusieurs articles, [GRISHMANN 1997], [RILOFF 1999], [SODERLAND 1999] ont montré l'intérêt d'appliquer des techniques d'apprentissage pour générer automatiquement des patterns à partir du texte. Cette approche semble intéressante même si, à notre connaissance, il n'existe pas encore d'applications industrielles.

Pour conclure, soulignons l'intérêt de notre approche basée sur les *Skill Cartridge*™ . Elles constituent à notre avis une technologie clé pour :

- L'intelligence économique
- La gestion des ressources humaines,
- La gestion de la connaissance et des savoir-faire,
- L'analyse de la relation client (CRM Customer Relationship Management)

Intégrer des connaissances sur le domaine d'application et les métiers est nécessaire pour que l'information extraite par les systèmes d'analyse ait une réelle valeur ajoutée. En effet, les consommateurs d'information ont des profils métiers différents (direction générale, marketing, recherche, ressources humaines,…) dans un ou plusieurs secteurs d'activité (bancaire, automobile, etc.). Chacun avec leurs besoins, leurs questions spécifiques, tous visant à anticiper les évolutions de l'environnement interne et externe de leur entreprise pour essayer de maintenir ou de développer sa compétitivité.

Les *Skill Cartridge*™ s'insèrent automatiquement dans une suite logicielle nommée *Insight Discoverer*™ *OnlineMiner*™ intégrant :

- un crawler de sites Web,
- un serveur de recherche documentaire incorporant un langage de requête de type SQL,
- un serveur d'extraction basé sur un analyseur morpho-syntaxique,
- un serveur de classification non hiérarchique et de catégorisation,
- des composants java de représentation graphique de tableaux et cartes.

OnlineMiner™ constitue un serveur d'analyse en ligne, accessible sur le Web, permettant d'organiser des documents collectés sur Internet par thèmes, en mode supervisé (catégorisation) ou non supervisé (classification) et de visualiser et synthétiser les résultats sous la forme de :

- tableaux synoptiques montrant, par exemple, l'évolution de l'opinion et de la satisfaction des clients (CRM) pour un produit donné, ou bien les actions d'une entreprise dans un domaine préétabli, …
- enquêtes historiques,
- projections et prévisions pour un produit, une marque, ou de manière générale, un objet donné (grâce à l'information temporelle extraite des textes),
- cartes thématiques,

L'interface de navigation hypertexte à partir de ces cartes et des tableaux synoptiques est organisée de manière permettre à l'analyste de l'information de confronter ses connaissances, ses besoins, ses centres d'intérêt personnels « Qui fait quoi, où, quand ? » avec les données initiales et leurs représentations par diverses méthodes d'analyse [GRIVEL 1997, 1999, 2000].

Bibliographie

[ABE 97] Abeillé A., Blache P. : Etat de l'Art : La Syntaxe - *TAL* 1997, vol.38, vol.2, pp.69-90.

[AUS 00]Aussenac-Gilles N. Biebow B. Szulman S. : Modélisation du domaine par une méthode fondée sur l'analyse de corpus IC'2000 - *Actes de la conférence Journées francophones d'Ingénierie des Connaissances Toulouse*, 10-12 mai 2000.

[BED 93] C. Bédécarrax, P. Coupet, P. Amsellem, C. Huot: Présentation de la station MARS de classification et d'analyse: application à l'analyse automatique de références bibliographiques de brevets extraites de WPI, Les systèmes d'informati on élaborée, *Congrès S.F.B.A.,* juin 1993.

[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 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.

[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 expérimentaitons, RFIA 2000, F. Schmitt et I Bloch Editeurs, Paris, France.

[GRISH 97]Grishman R. Yangarber R. Issues in Corpus Trained Information ExtractionTeresa Pazienza, Springer Notes in Artificial Intelligence, Springer-Verlag, 1997

[GRI 99] Grivel L. : 'HENOCH, un outil d'analyse de corpus d'information scientifique et technique'*, Le Micro Bulletin Thématique du CNRS* n°3, L'information scientifique et technique et l'outil Internet, CNRS-DSI, p.27-44, 1999.

[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.

[HAB 98] Benoît Habert, Adeline Nazarenko, Pierre Zweigenbaum, and Jacques Bouaud. Extending an existing specialized semantic lexicon. In Antonio Rubio, Navidad Gallardo, Rosa Castro, and Antonio Tejada, editors, *First International Conference on Language Resources and Evaluation*, pages 663-668, Granada, 1998.

[HOB 97]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 Processin*g. Cambridge MA: MIT Press. . (1997)

[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.