Une Application Industrielle d'Extraction de l'Information pour l'Intelligence Economique

Bianka Bushbeck, Luc Grivel, Sylvie Guillemin-Lanne, Christian Lautier

Une Application Industrielle d'Extraction de l'Information pour l'Intelligence Economique

Buschbeck Bianka — Grivel Luc — Guillemin-Lanne Sylvie — Lautier Christian

TEMIS Text Mining Solutions

59, rue de Ponthieu
75 008 Paris

{bianka.buschbeck, luc.grivel, sylvie.guillemin–lanne, christian.lautier }
@temis-group.com

RÉSUMÉ. *Cet article développe une méthodologie de construction de règles d'extraction d'information et présente des exemples de résultats pour une application industrielle en veille concurrentielle. L'article met l'accent sur les avantages que représente l'organisation des règles d'extraction en niveaux hiérarchiques. Celle-ci permet de gérer des hiérarchies de patterns et de contrôler leur exécution sur les corpus. Le but est de fournir aux PMI locales des informations marketing sur les produits, les concurrents, leurs actions de communication (annonce, intention, rumeur,...), leurs transactions (achat, acquisition, fusion, cession, vente,...) dans leur domaine d'activité (business général, agroalimentaire, artisanat et tourisme).*

ABSTRACT .*This article focuses on the advantages of organizing rules for knowledge extraction in a hierarchical order. It details the methodology to develop extraction rules, and illustrates the results with examples taken from real-life competitive intelligence applications.*

MOTS-CLÉS : *fouille de données textuelles, ingénierie des connaissances, extraction d'information, règle d'extraction.*

KEYWORDS: *text mining, knowledge engineering, knowledge extraction, information extraction, extraction rule.*

ECG 2002 Actes Communications d'entreprises
Nom de la revue. Volume X – n° X/2001, pages 1 à X

1. Un modèle d'extraction d'information appliqué à l'intelligence économique

1.1. *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"* [WIL 97]. En d'autres termes, l'extraction d'information est le processus qui permet d'identifier l'information pertinente, où les critères de pertinence sont définis sous forme de patrons (templates) à remplir. Par exemple, pour une application d'intelligence économique:

– 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 ?

Considérons l'exemple d'extraction ci-dessous, où l'objectif est de lier les acteurs, des compagnies de l'agroalimentaire, aux actions décrites (un rachat d'entreprises).

In February 2000 Bestfoods bought private Brazilian food giant Arisco Produtos Alimenticios and its 750 products (including condiments, soups, and seasonings) for $490 million.

In February 2000 Bestfoods bought private Brazilian food giant Arisco Produtos Alimenticios for $490 million.

when: In February 2000

who:
/organisation Bestfoods

which_information:
/Competitive Intelligence/**buying acquisition** bought

whom:
/actor qualifier private Brazilian food giant
/potential company Arisco Produtos Alimenticios

/financial operation
/Money/Dollar for $490 million

Figure 1. *Extraction d'information : Exemple de rachat d'une entreprise dans l'agroalimentaire.*

L'information extraite est définie dans des rôles (*who, whom*, pour les acteurs, *which_information* pour la nature de l'action*, when* pour la date), qui s'expriment sous forme d'attribut valeur et peuvent ainsi être capturés automatiquement.

Avant de présenter la méthodologie d'extraction, considérons les types d'actions à représenter dans une application dédiée à l'intelligence économique.

1.2. *Les actions de l'intelligence économique(IE)*

Les actions pertinentes en IE concernent d'une part les actions de communication (annonce, intention, rumeur,...) et les transactions proprement dites (achat, acquisition, fusion, cession, vente,...).

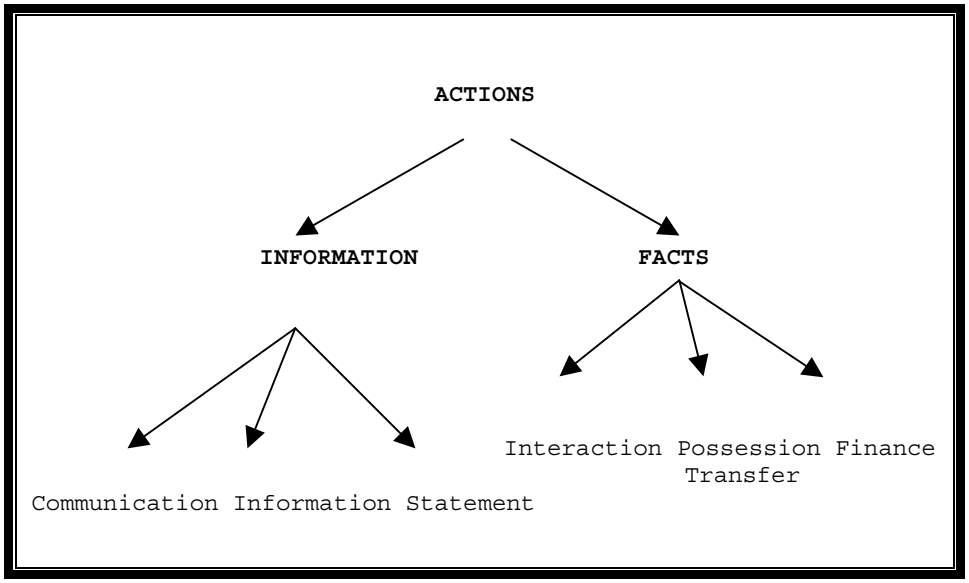


Figure 2. *Description des actions en Intelligence Economique*

Toutes les actions susceptibles d'apporter une information stratégique sur le domaine économique étudié font donc l'objet d'une description, comme le suggère le tableau ci-dessous :

Interaction			*Possession Transfer*	*Finance*	
	Alliances	*Divorce*			
transaction	agreement	divorce	possession	compensation	valorisation
bid	partnership		buying_acquisit	investment	depreciation
	control		selling_cession	asset	evaluation
				charge	
				gain	
				loss	

Figure 3. *Organisation de l'information pertinente en Intelligence Economique*

Les faits identifiés peuvent être annoncés bien qu'ils n'aient pas encore eu lieu. Il peut s 'agir d'une rumeur, d'une annonce ou d'une intention. Leur extraction donne ainsi à l'utilisateur une meilleure compréhension de l'information.

Communication			*Think*		
Information	*Communication*	*Statement*	*Wonder*	*Judge_estimate*	*Plan*
announcement	approval	conclusion	wonder	expectation	intention
declaration	confirmation			decision	
explanation	refusal				
	proposal				
	negociation				

Figure 4. *Organisation de l'information : L'acte de communication*

2. L'approche TEMIS

2.1. La notion de Skill Cartridge™

Une *Skill Cartridge*™ ou cartouche de connaissance est une hiérarchie de composants de connaissance décrivant l'information à extraire. 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.

Les Skill Cartridges™ sont construites à partir d'un corpus d'entraînement. Notre approche est fondée sur l'utilisation de bases terminologiques existantes et sur l'acquisition de connaissances à partir du corpus selon un processus itératif. Elle s'inscrit dans le courant BCT, bases de connaissances terminologiques (groupe TIA, Groupe Ingénierie des Connaissances GRACQ) [ABE 97], [AUS 00], [BOU 98, 99, 00], [CHA 00], [CON 00], [HAB 98], [NAZ 97], [JAC 94].

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, télécom.etc)

– le regroupement des termes ainsi définis sous des descripteurs sémantiques eux-mêmes organisés, selon les besoins, en une hiérarchie simple [GRI 01a],

– la définition des règles d'extraction d'information s'appliquant aux concepts de premier niveau (un terme renvoie à un descripteur) , afin de dégager des concepts de niveau supérieur et d'aboutir aux concepts visés, comme nous le verrons en seconde partie de notre exposé,

– l'exécution interactive des règles d'extraction sur le corpus d'entraînement afin d'évaluer le résultat de l'extraction et de valider les composants pour un domaine d'application.

2.2 La hiérarchisation de l'information en niveaux

Le module d'extraction utilise les cartouches de connaissances qui définissent par un ensemble d'expressions quels concepts déclencher. Il prend en entrée un texte étiqueté par une analyse morpho-syntaxique et évalue de gauche à droite l'ensemble des expressions en ne gardant que les plus longues.

Pour gérer des priorités différentes que « le plus à gauche, le plus long », chaque cartouche peut être décomposée en niveaux contenant chacun un sous-ensemble d'expressions. Chaque concept trouvé à un niveau remplace les unités qui l'ont déclenché, permettant aux niveaux supérieurs d'utiliser ce concept pour bâtir d'autre concepts.

Pour optimiser la vitesse et la taille des dictionnaire, le module d'extraction utilise la technologie des transducteurs [HOB 97], [KAR 00].

Le formalisme des expressions est un sur-ensemble des expressions régulières combinant l'accès aux formes de surface, aux tags grammaticaux et aux lemmes déclenchant des concepts qui peuvent être réutilisés pour définir d'autres concepts.

L'idée de base est de construire des patterns. Un pattern est une expression régulière qui identifie le contexte de syntagmes pertinents et les délimiteurs de ces syntagmes. Par exemple, les expressions de temps seront regroupées sous le descripteur When :

```
In February 2000

<concept name="~month" language="english">
    February
</concept>

<concept name="~date" language="english">
       [12][0-9][0-9][0-9]
</concept>

<concept name="When">
       ~PREP / ~month / ~date
</concept>

when:  in February 2000
```

Figure 5..*Règles de construction d'une expression de temps*

En associant un concept 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 de niveau supérieur.

```
for $490 million

<concept name="~Money" >
  <concept name="~Dollar" >
      ~$[0-9]+ / ~million
  </concept>
</concept>

<concept name="~financial_operation" >
      ~PREP / ~Money
</concept>

~financial_operation:  for $490 million
      ~Money/Dollar        $490 million
```

Figure 6. *Règle de construction d'une expression financière*

La hiérarchisation de l'information par niveaux permet ainsi de gérer des hiérarchies de patterns et de contrôler leur exécution sur les corpus. Toute séquence de mot regroupée au sein d'un pattern, l'est de façon définitive pour tous les niveaux suivants et ne pourra être disloquée pour construire un pattern concurrent à un niveau supérieur. Un mot isolé ne gardera pas l'étiquette sémantique qui lui a été attribuée à un niveau inférieur, sauf si l'option « nextlevel=true » a été activée. Dans le cas contraire, il perd son étiquette sémantique en changeant de niveau et redevient disponible pour la construction d'un autre pattern.

Sur le plan méthodologique nous préconisons de décrire :

au niveau 0 :

– les règles de reconnaissance des patterns de lieu, de valeurs numériques, monétaires, d'expressions temporelles
– les données relatives au domaine d'activité décrit. Les acteurs du domaine (question who) ou les objets du domaine (question what), exprimés sous forme de liste et organisés sous des concepts « descripteurs ».

au niveau 1 :

– les règles de construction «builder rules» des concepts dits intermédiaires, eux-mêmes utilisés dans des concepts de niveau supérieur. Dans notre exemple, il s'agit des concepts relatifs à l'Intelligence économique, tels que finance_amount, ou stake ou customers par exemple.

au niveau 2 :
– les règles dites «guesser rules» qui vont construire des concepts « potentiels » qui ne deviendront définitifs qu'au moment de leur appel dans une règle à un niveau supérieur. Le but visé ici, est de garantir la généricité de la cartouche.

Dans l'exemple d'Intelligence économique que nous avons choisi d'illustrer, il s'agit d'étendre la reconnaissance des acteurs, noms de compagnies ou expressions qualifiantes aux autres noms que ceux prévues dans les listes définies au niveau 0. Le pattern ainsi étiqueté «potential company» ne sera effectif que lors de l'application des règles de niveau supérieur.

```
<concept name="~GroupWord" >
     Ltd
</concept>

<concept name="~potential_company" >
  (~Lieu)? / (#NP)+ / ((in)? / ~Lieu)? / (~GroupWord)*
</concept>

potential company: Arisco Produtos Alimenticios
```

Figure 7. *Extraction d'information : règle de niveau 2*

au niveau 3 :

– les règles dites «linker rules» qui lient les concepts entre eux pour construire le pattern visé. A ce niveau, intervient un opérateur *ANY* qui matche toute séquence de mots jusqu'au prochain concept rencontré. L'exemple présenté en première partie tient lieu d'illustration de l'application de la règle d'extraction exposée ci-dessous :

```
<concept name="~Extraction_for_Competitive_Intelligence" >

(~When / (~Lieu)* / ANY)? /

{who: (~actor)} / ANY / (~When / ANY)?

{which_information:
((~Announcement|~Rumor|~financial_operation) / ANY)*
~Competitive_Intelligence } / ANY /

{whom: (~actor)} /

(ANY / ~financial_operation)* / (ANY / ~When)?

</concept>
```

Figure 8. *Extraction d'information : règle de niveau 3*

Voyons à présent, niveau par niveau, comment s'est construit le pattern extrait :

Au niveau 0 ont été reconnus les concepts :

/When	{in February 2000},
/Organisation	{Bestfoods},
/CollectiveWord	{giant},
/Money/Dollar	{$490 million}

Au niveau 1 ont été reconnus les concepts :

/Competitive Intelligence/buying acquisition	{buy} ,
/financial operation	{for $490 million}

Au niveau 2 ont été reconnus les concepts :

/actor qualifier	{private Brazilian food giant}
/potential company	{Arisco Produtos Alimenticios}

Les concepts ainsi identifiés vont alors être liés entre eux par les règles de niveau 3 :

```
Level 2: Input for level 3

      /When                                    {in February 2000}
      /Organisation                                 {Bestfoods}
      /Competitive Intelligence/buying acquisition       {buy}
      /actor qualifier              {private Brazilian food giant}
      /potential company            {Arisco Produtos Alimenticios}
       and                  and          CC
       its                  it           PP$
       750                  750          CD
       products             product      NNS
       (                    (            PUNCT
       including            include      BG
       condiments           condiment    NNS
       ,                    ,            CM
       soups                soup         NNS
       ,                    ,            CM
       and                  and          CC
       seasonings           seasoning    NNS
       )                    )            PUNCT
       for                  for          IN
      /financial operation                   {for $490 million}
         /Money/Dollar                           {$490 million}
```

Figure 9. *Visualisation des patterns identifiés*[1]

3. Une application industrielle

Une application industrielle a été menée avec Telcal/Intersiel et IBM Italie**.** Telcal (Telematica Calabria) est un consortium créé à partir de Finsiel et Telecom Italia, dont l'objectif est de participer au développement économique de la province de Calabre en Italie du Sud. Il s'agit d'un système de surveillance de la concurrence dans les domaines de l'agriculture, de l'artisanat et du tourisme. Le but est de fournir aux PMI locales des informations marketing sur les produits, les concurrents et leurs actions dans leur domaine d'activité (business général, agroalimentaire, artisanat et tourisme). Une application a été développée et couple les Skill Cartridges et Insight Discoverer Extractor TEMIS [ZAN 01] [GRI 01b].

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

[1] le format des patterns reconnus s'écrit : /pattern {expression}
le format des mots non extraits, identique au format d'entrée avant l'extraction, reporte en col.1 la forme du mot, en col.2 le lemme correspondant, en col.3 le tag grammatical.

OnlineMiner™ est un serveur accessible sur le Web permettant de collecter et d'organiser des documents par thèmes, en mode supervisé (catégorisation) ou non supervisé (classification) et de visualiser et synthétiser les résultats sous la forme de tableaux, graphiques ou cartes. Plusieurs approches peuvent être combinées selon l'objectif recherché :

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 des concepts

Une telle combinaison permet de réduire le temps passé par l'utilisateur pour trouver et analyser les documents qui l'intéressent lorsque une requête ramène beaucoup de documents (comme par exemple la requête 'food%' ci-dessous).

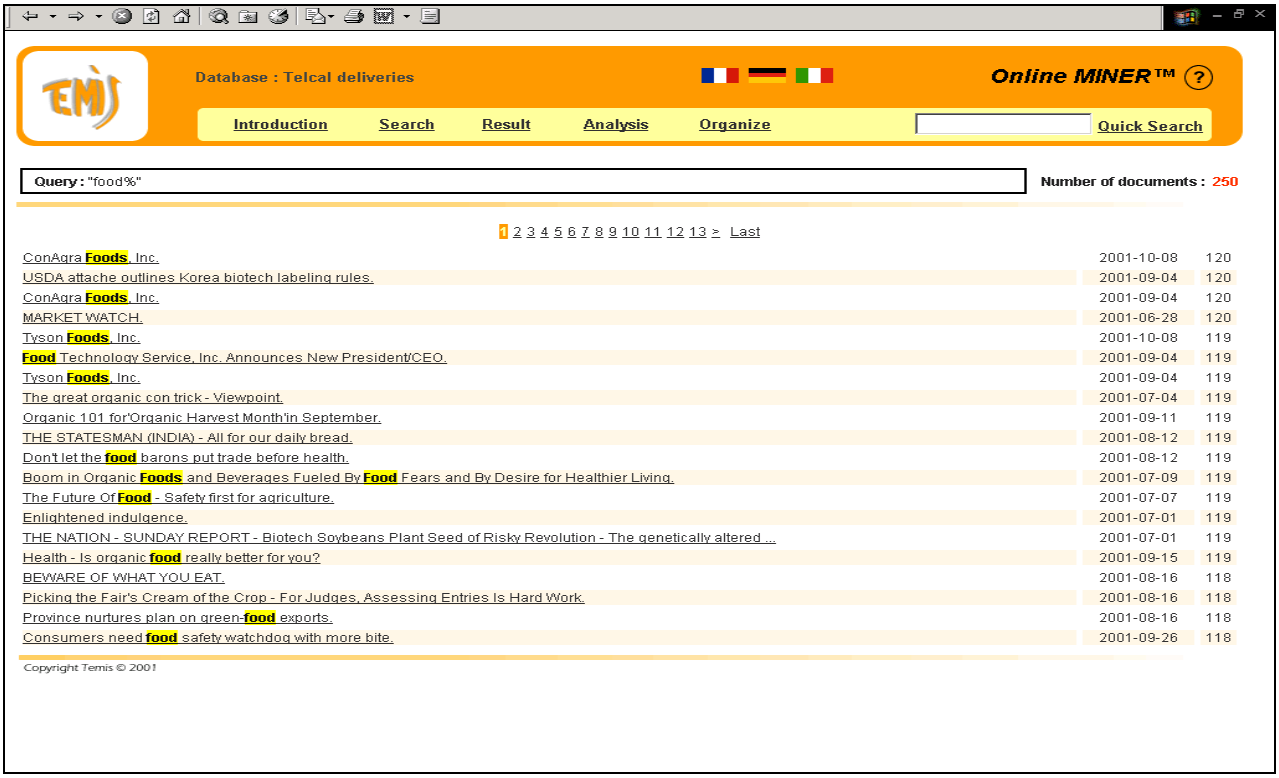


Figure 10. *OnlineMiner*™ *: visualisation des résultats à partir d'une requête*

Il est par exemple possible de visualiser la liste des concepts extraits à partir des textes sélectionnés ci-dessus, ainsi que les documents concernés.

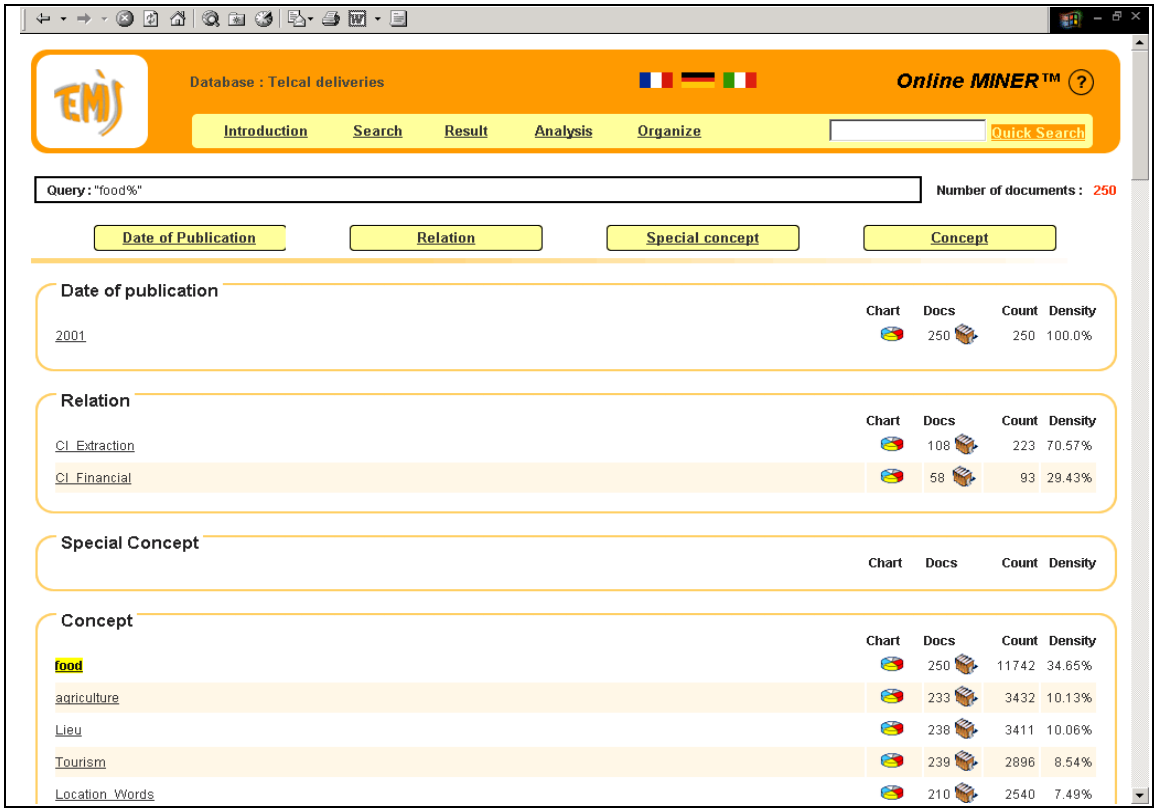
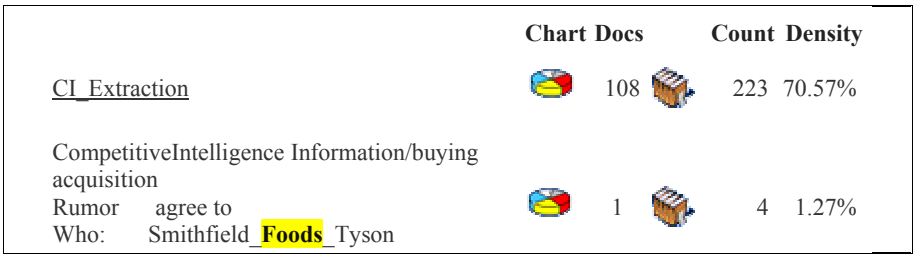


Figure 11. *OnlineMiner™ : Analyse des résultats et liste des concepts extraits*

Puis d'appréhender les relations qui ont été extraites :

CompetitiveIntelligence Information/buying acquisition,
CompetitiveIntelligence Information/stakes acquisition.



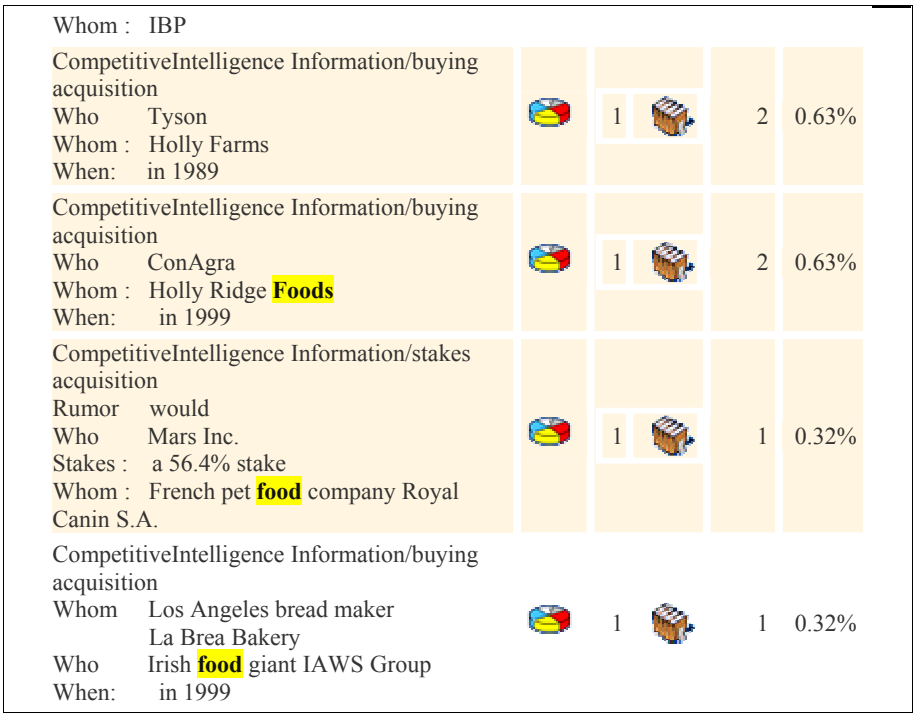


Figure 12. *OnlineMiner™ : Analyse des résultats et visualisation des relations extraites*

4. Conclusion

Nous avons montré les avantages d'une organisation de règles d'extraction en niveaux hiérarchiques. Elle permet de gérer des hiérarchies de patterns et de contrôler leur exécution sur les corpus. La méthodologie de construction de règles d'extraction décrite dans cet article a été rapidement maîtrisée par les linguistes. Nous envisageons, à présent, de concentrer nos efforts pour généraliser les composants de connaissance dans un sens qui nous permette de les réutiliser sur plusieurs domaines et/ou de les customiser aux applications clients. Les résultats de ce projet sont très prometteurs.

Par ailleurs, un programme de recherche est en cours pour aider les linguistes dans leur tâche. L'objectif est de développer un environnement logiciel pour aider à la customisation des Skill Cartridge™. Pour ce projet qui devrait se terminer en 2002, TEMIS s'est vu décerner récemment le label technologie-clé par l'**ANVAR**. Aujourd'hui, TEMIS peut livrer des Skill Cartridges™ en français et en anglais. Nous envisageons, dans un futur proche, de développer des Skill Cartridges™ en italien, espagnol, allemand.

Références

[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*. mai 2000.

[BAS 00] Basili, R., Pazienza, M. T., and Vingini, M. (2000). Corpus driven learning of event recognition rules. In Machine learning for Information Extraction Workshop, Berlin, Allemagne.

[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

[BRI 94] Brill, E. and Resnik, P. (1994). A rule based approach to prepositional phrase attachment disambiguation. In Proceedings of the 15th annual conference on Computational Linguistics.

[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.

[COW 96] J. Cowie, & W. Lehnert (1996) Information Extraction, in (Y. Wilks, ed.) Special NLP Issue of the Comm. ACM.

[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] Faure D. 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.

[FAU 98] Faure, D. and N'edellec, C. (1998a). A Corpus based Conceptual Clustering Method for Verb Frames and Ontology Acquisition. In Velardi, P., editor, LREC workshop on Adapting lexical and corpus ressources to sublanguages and applications, pages 5--12, Granada, Espagne.

[GRISH 97] Grishman, R. (1997). 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. 1418.

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

[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 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 00] Grivel L. : L'hypertexte comme mode d'exploitation des résultats d'outils et méthodes d'analyse de l'information scientifique et technique, Phd thesis, Université Aix-Marseille III,. 10 janvier 2000.

[GRI 01a] 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 », 3ème congrès du Chapitre français de l'ISKO International Society for Knowledge Organization, 5-6 juillet 2001

[GRI 01b] Grivel Luc, Guillemin-Lanne Coupet, Pascal, Huot Charles « Analyse en ligne de l'information : une approche permettant l'extraction d'informations stratégiques basée sur la construction de composants de connaissance » VSST 2001 Veille Scientifique et Stratégique, Barcelone, 15-19 octobre 2001.

[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.

[HAR 88] Harris, Z. (1988). Language and Information. Columbia University Press, New York.

[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)

[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)

[JAC 97] Jacquemin, C. (1997). Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. PhD thesis, Universit'e de Nantes.

[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*).*

[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 93] E. Riloff and W. Lehnert. Automated Dictionary Construction for Information Extraction from Text, In *Proceedings of the Ninth IEEE Conference on Artificial Intelligence for Applications,* pages. 93-99. IEEE Computer Society Press. 1993.

[RIL 95] E. Riloff, and J. Shoen (1995) Automatically acquiring conceptual patterns without an annotated corpus, Proc. Third Workshop on Very Large Corpora.

[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.

[WIL 97] Wilks, Y. (1997). Information Extraction as a Core Language Technology. In Pazienza, M. T., editor, Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, Frascati, Italie. LNAI Tutorial, Springer. 1418.

[ZAN 01] Zanasi, A. Text Mining : The New Competitive Intelligence Frontier. Real Application Cases in Industrial, Banking and Telecom/SMEs World» VSST 2001 Veille Scientifique et Stratégique, Barcelone, 15-19 octobre 2001.

[ZWE 00] Zweigenbaum, P. and Grabar, N. (2000). Expériences d'acquisition automatique de connaissances morphologiques par amorçage à partir d'un thesaurus. In Schmitt, F. and Bloch, I., editors, 12eme Congrès Francophone AFRIF AFIA de Reconnaissance des Formes et Intelligence Artificielle (RFIA'2000), volume II, pages 101--110, Paris, France.