Une méthodologie et un environnement d'aide à la construction de composants de connaissance pour l'Extraction d'Information

Christophe Aubry, Luc Grivel, Sylvie Guillemin-Lanne, Christian Lautier

Une méthodologie et un environnement d'aide à la construction de composants de connaissance pour l'Extraction d'Information

Aubry Christophe, Grivel Luc, Guillemin-Lanne Sylvie, Lautier Christian

TEMIS Text Mining Solutions

59, rue de Ponthieu
75 008 Paris

{christophe.aubry, 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.

Mots Clés

fouille de données textuelles, ingénierie des connaissances, extraction d'information, règle d'extraction.

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.

Keywords

text mining, knowledge engineering, knowledge extraction, information extraction, extraction rule.

1. Introduction

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. Elle est par nature excessivement hétérogène (structure, sémantique, formats, etc) . De tels volumes, une telle hétérogénéité exigent de nouveaux outils capables d'analyser et de structurer des documents textuels en prenant en compte le sens de ces documents.

La société TEMIS s'est fixée pour objectif le développement de systèmes d'analyse de données textuelles (text mining) basés sur la connaissance. De tels systèmes s'appliquent aux données, produites ou reçues, par une société dans son domaine pour en extraire les informations pertinentes selon divers points de vue tels que:

- L'analyse concurrentielle (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)

Un des enjeux essentiels de cette activité est la capacité à développer des systèmes pouvant s'adapter à de nouveaux domaines d'application, de nouveaux secteurs d'activité, de nouvelles thématiques d'analyse ou à une autre langue [POIBEAU 2002],[GRISH 97].

Dans le cas d'une thématique d'analyse donnée (comme l'intelligence économique), on observe que le coût de la personnalisation des systèmes d'extraction est particulièrement élevé si les modifications doivent être effectuées par les développeurs des règles d'extraction (le plus souvent des linguistes). Les 'clients industriels' de ces systèmes d'extraction tiennent à garder le 'contrôle' de leurs ressources et à capitaliser/valoriser les connaissances qu'ils ont de leur domaine et de leur

environnement (concurrentiel, scientifique, technique, juridique, …).

Pour intégrer ces données spécifiques (sociétés ou acteurs du domaine, produits du domaine, actions propres au domaine,...) à une thématique d'analyse (Intelligence économique, analyse de courriels, étude de CV), nous avons développé un formalisme et une méthode permettant d'exprimer et de construire des règles[1] d'extraction génériques et de définir une structure où ces dictionnaires peuvent facilement s'insérer.

Après un court rappel sur l'extraction d'information, nous décrirons, notre approche. Notre exposé sera illustré d'exemples concrets extraits d'applications réelles d'intelligence économique. Nous décrirons également la première version opérationnelle d'un environnement permettant de mettre en œuvre cette méthode.

L'objectif, à terme, est d'améliorer les conditions d'usage du système d'extraction de manière à ce que l'effort nécessaire pour sa personnalisation par des experts (du domaine d'application), le plus souvent ni informaticien ni linguiste, soit 'raisonnable' (c'est-à-dire à sa portée en un temps limité de formation).

2. L'approche TEMIS

2.1 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. Les critères de pertinence sont définis sous forme de patrons d'extraction et les données extraites sont transposées dans des formulaires (templates) prêts à être remplis.

Pour une application d'intelligence économique, on cherchera à décrire:

– 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, extrait d'une application d'intelligence économique où l'objectif est de lier des acteurs du domaine, en l'occurrence 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 exprimée sous forme de couples attribut-valeur (rôles) : *who, whom*, pour les acteurs, *which_information* pour la nature de l'action, *when* pour la date.

2.2 Le formalisme

L'objectif est de construire des patrons d'extraction *(extraction patterns)* [GRISH 97]. Un patron d'extraction décrit une structure syntaxique de surface comportant des éléments lexicaux, des tags grammaticaux et des éléments typés sémantiquement [POI 02]. En d'autres termes, un patron d'extraction est une expression régulière qui identifie le contexte de syntagmes pertinents et les délimiteurs de ces syntagmes.

Notre formalisme permet de décrire des règles d'extraction par 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.

Dans l'exemple ci-dessous, les expressions de temps sont regroupées sous le 'concept' *When* :

[1] Notre expérience avec nos clients est qu'ils n'ont pas de difficulté à lire les règles d'extraction dans le formalisme que nous proposons, mais par contre qu'il leur serait difficile de les écrire directement.

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 2..*Règle de construction d'une expression de temps*

En associant un concept à un pattern, une règle ajoute de l'information à une séquence de mots, par exemple, en lui attribuant un nom de classe sémantique qui peut être ensuite utilisé dans d'autres règles.

3. Skill Cartridge™ et Méthodologie

3.1 Structure modulaire

Notre approche de développement des composants de connaissance est guidée par l'idée de favoriser la réutilisation de composants, de la même manière que dans un langage de programmation, il est possible de définir des classes d'objets et de les utiliser.

Nous modélisons et organisons l'information à extraire selon une hiérarchie de composants de connaissance. Cette hiérarchie est appelée ***Skill Cartridge™*** ou cartouche de connaissance. Un composant de connaissance peut avoir la forme d'un dictionnaire ou d'un ensemble de règles d'extraction.

Pour développer des composants de connaissance modulaires pour différents domaines d'activité et/ou langues, il faut une méthodologie pour organiser les données du/des dictionnaires et assurer la généricité des règles d'extraction.

3.2 La hiérarchisation de l'information en niveaux

Notre système utilise la règle classique « le plus à gauche, le plus long » pour savoir quelle séquence de mots associer aux patterns candidats. Afin de gérer des priorités différentes, une cartouche peut être décomposée en niveaux contenant chacun un sous-ensemble d'expressions. Un concept extrait à un niveau donné encapsule les unités qui l'ont déclenché, rendant celles-ci inaccessibles aux niveaux supérieurs et permettant d'utiliser ce concept pour en bâtir d'autres.

La hiérarchisation de l'information par niveaux permet ainsi de gérer des hiérarchies de patterns et de contrôler leur exécution sur les corpus. Toute séquence de mots regroupée au sein d'un pattern, l'est de façon définitive pour tous les niveaux suivants et ne pourra donc pas être disloquée pour construire un pattern concurrent à un niveau supérieur. Un mot isolé, qui n'a pas participé à la construction d'un pattern perd, en changeant de niveau, son étiquette sémantique et redevient disponible pour la construction d'un autre pattern.

Pour la thématique d'intelligence économique, nous avons choisi de décrire :

au niveau 0 : les dictionnaires spécifiques et les expressions qui restent valides tout au long du processus d'extraction, par exemple,

- les règles de reconnaissance des expressions de lieu, de valeurs numériques, monétaires, d'expressions temporelles,

- les acteurs du domaine (question *who*) ou les objets du domaine (question *what*), exprimés sous forme de liste et organisés sous des concepts « descripteurs ».

```
<concept name="~CompanyName" >
    <concept name="~Telecom_Operator">

        <concept name="~FranceTelecom_Operator">
            France / Télécom
            France / Telecom
            Itinéris
            ...
```

Figure 3: *Description des acteurs du domaine* [2]

au niveau 1 : les règles de construction «*builder rules*» des concepts intermédiaires de la thématique qui seront utilisés dans des concepts de niveau supérieur.

Dans notre exemple, il s'agit de concepts relatifs à l'Intelligence économique, tels que ~finance_amount qui décrit un revenu, ~stake ou ~customers par exemple.

```
for $490 million
```

[2] Le concept ~CompanyName permet d'intégrer un dictionnaire spécifique, ici des noms d'acteurs du domaine des télécommunications (~Telecom_Operator et ~FranceTelecom_Operator). Ainsi ~CompanyName peut être appelé dans des règles générales indépendantes du domaine d'application.

```
<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 4. *Règle de construction d'une expression financière*

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, tout en évitant de maintenir des listes trop lourdes à gérer.

Dans l'exemple d'Intelligence économique que nous avons choisi d'illustrer, il s'agit donc d'étendre la reconnaissance des acteurs, noms de compagnies ou expressions qualifiantes aux autres noms ou séquences de noms que ceux prévus 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 5. *Extraction d'information : règle de niveau 2*

au niveau 3 : les règles dites «*linker rules*» qui lient les concepts entre eux pour remplir le formulaire 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 6. *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
/financial operation	{for $490 million}	
/Money/Dollar	{$490 million}	

Figure 7. *Visualisation des patterns identifiés*[3]

L'ensemble des extractions peut ainsi être affecté à des formulaires (par thème, par acteur, etc.) Cf Figure 8 en annexe.

4. Skill Cartridge™ Studio

4.1 Design

Notre approche générale est fondée sur l'utilisation de bases terminologiques et/ou ressources existantes[4] et sur l'acquisition de connaissances à partir de corpus selon un processus itératif.

Les Skill Cartridges™ sont en effet construites à partir de corpus. Les étapes de construction sont :

– l'analyse morpho-syntaxique des textes constituant le corpus puis la définition du vocabulaire pertinent relatif au secteur d'activité ou au domaine économique étudié (agroalimentaire, automobile, télécommunication, etc)

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

– 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, ,

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

Cette approche s'inscrit dans le courant BCT, bases de connaissances terminologiques (Groupe TIA, Groupe Ingénierie des Connaissances GRACQ) [BOU 98, 99, 00], [HAB 98], [NAZ 97], [JAC 97], [TOU98].

Notre objectif est de développer un environnement (Skill Cartridge™ Studio) permettant de mettre en œuvre la méthodologie décrite plus haut, mais aussi d'intégrer et d'évaluer diverses méthodes/stratégies d'acquisition automatique ou semi-automatique de données (ressources, classe sémantiques, patrons syntaxiques) [POIBEAU 02], [RIL99], [SDO99], [BOU 99], [TOU98], [ZWE00].

Au stade actuel de développement, seul le premier objectif (formalisme vs méthodologie) est réalisé. Toutefois, Skill Cartridge™ Studio (SCS) est basé sur une architecture ouverte permettant d'intégrer/plugger facilement des composants logiciels et donc de nouvelles fonctionnalités.

Cette architecture, issue du projet Eclipse - développé par IBM pour gérer des projets écrits en JAVA -, permet de définir toutes les ressources relatives à un projet et d'offrir des vues spécifiques pour un type de ressource donné. Par exemple, pour la ressource Skill Cartridge™, définir une vue permettant de rechercher, créer, éditer une Skill Cartridge™).

4.2 L'environnement Skill Cartridge™ Studio

Nous nous limiterons ici à décrire quelques unes de ces vues pour la ressource Skill Cartridge™, dont l'ensemble des composants est au format XML (cf. les exemples fournis dans cet article).

- Edition

La vue 'Edition' (Figure 9 en annexe) guide le linguiste dans sa tâche de construction des Skill Cartridges™,. Elle lui permet de choisir les composants de base pour démarrer un nouveau projet en naviguant dans la hiérarchie de la Skill Cartridge™, de naviguer dans les concepts décrits dans un composant, de définir une nouvelle règle d'extraction avec une assistance dans la gestion des attributs et de leurs valeurs, conformément à la DTD décrivant la syntaxe des règles d'extraction.

- Compilation

Skill Cartridge™ Studio intègre un compilateur de Skill Cartridge™ qui procède à la vérification syntaxique des composants, et traduit les dictionnaires et règles d'extraction en un ensemble de transducteurs[HOB 97], [KAR 00].

[3] 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.

[4] bases clients, WORDNET, etc.

La vue de compilation permet de sélectionner les options de compilation : choisir la (les) langue[5] (s) de la Skill Cartridge™, définir le niveau de détail des messages (debugging), en liant le message d'erreur avec la règle qui en est la cause.

- Test

Skill Cartridge™ Studio communique avec un serveur d'extraction qui lit les cartouches de connaissances et extrait les concepts à déclencher.

La vue de test permet de visualiser les résultats de l'extraction , par sur-brillance des concepts extraits sur le texte et par une visualisation de la correspondance règles vs concepts extraits.

5. Conclusion

Le formalisme décrit dans cet article permet de développer des systèmes d'extraction d'information **personnalisés** et **évolutifs** prenant en compte des besoins d'analyse divers de fonctions (commercial, recherche, technique) et profils métiers variés (automobile, banque, etc.).

Par exemple, un système de surveillance de la concurrence dans les domaines de l'agriculture, de l'artisanat et du tourisme basé sur le système d'extraction décrit dans cet article a été développé en collaboration Telcal/Intersiel[6] et IBM Italie[ZAN 01], [GRI 01b]**.** Le système fournit 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).

Le développement de cette première version opérationnelle de la Skill Cartridge™ Studio est terminé depuis le mois de mars 2002. Celle-ci est utilisée actuellement par des stagiaires de **DESS** *Traitement et Valorisation de l'Information Textuelle* de **l'Université de Poitiers** et par une stagiaire du **DESS** *Ingéniérie de la Langue et Société de l'Information* de **l'Université de Paris-Sorbonne**. Il est donc encore trop tôt pour tirer les premières leçon concernant l'interface.

Sur le plan méthodologique, nous avons montré les avantages d'une organisation de règles d'extraction par niveaux. Cette organisation 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 ces stagiaires, maîtrise qui s'est traduite immédiatement par une augmentation de leur productivité lors de l'écriture de Skill Cartridges™.

Références

[AIT 97] Aït-Mokhtar S., Chanod J.-P. : Incremental finite-state parsing. *In proceedings of the 5th International Conference on Applied Natural Language Processing* (ANLP'1997), Washington, 1997, pp. 72-79.

[APP 95] Appelt D., Hobbs J., Bear J., Israel D., Kameyama M., Kehler A., Martin D., Myers K. et Tyson M. SRI International Fastus System : MUC-6 test results and analysis. I*n proceedings of the 6th Message Understanding Conference* (MUC-6), Morgan Kaufmann, San Francisco, 1995, pp. 237-248.

[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 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 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.

[BUS 02] Bushbek Bianka, Grivel Luc, Guillemin-Lanne Lautier Christian « Une application industrielle d'extraction d'informations pour l'Intelligence Economique » EGC 2002 Extraction et Gestion des Connaissance, Montpellier, 21-23 janvier 2002.

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

[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.

[5] actuellement, en français, en anglais, en italien, espagnol, et en allemand.

[6] 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.

[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.

[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 Sylvie, 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.

[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 Processing*. Cambridge MA: MIT Press. (1997)

[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.*.

[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.

[POI 02] Poibeau T. : Extraction d'information à base de connaissances hybrides, Thèse, Université Paris-Nord, 8 mars 2002.

[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.

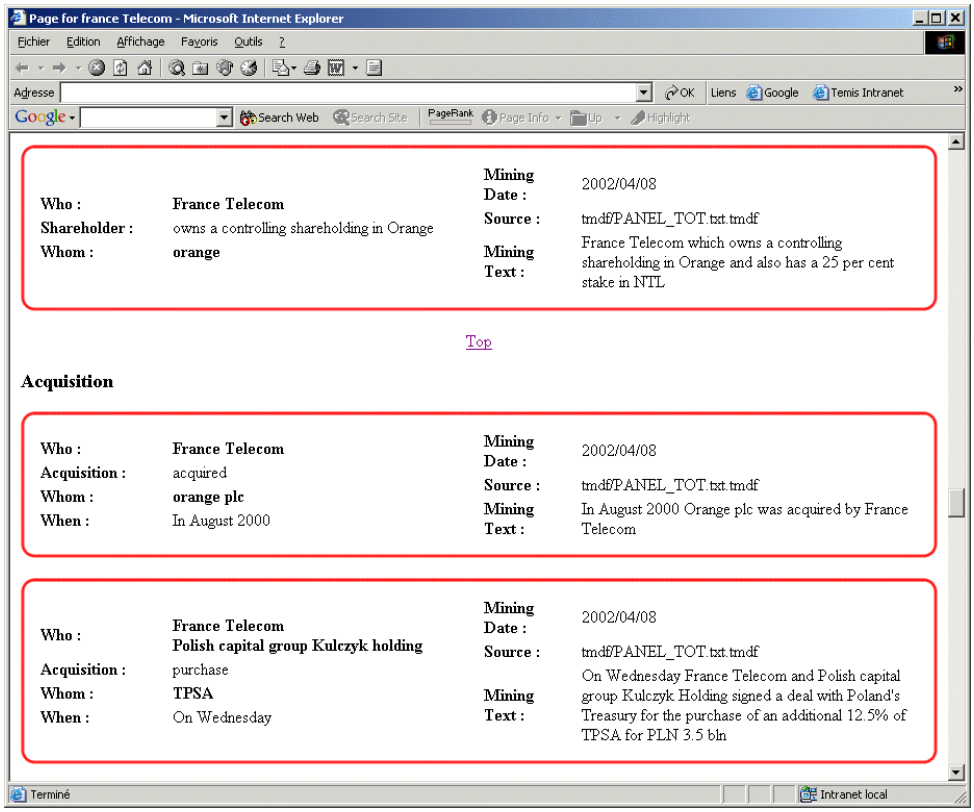

Figure 8. Exemple d'une fiche entreprise regroupant des extractions en Intelligence Economique

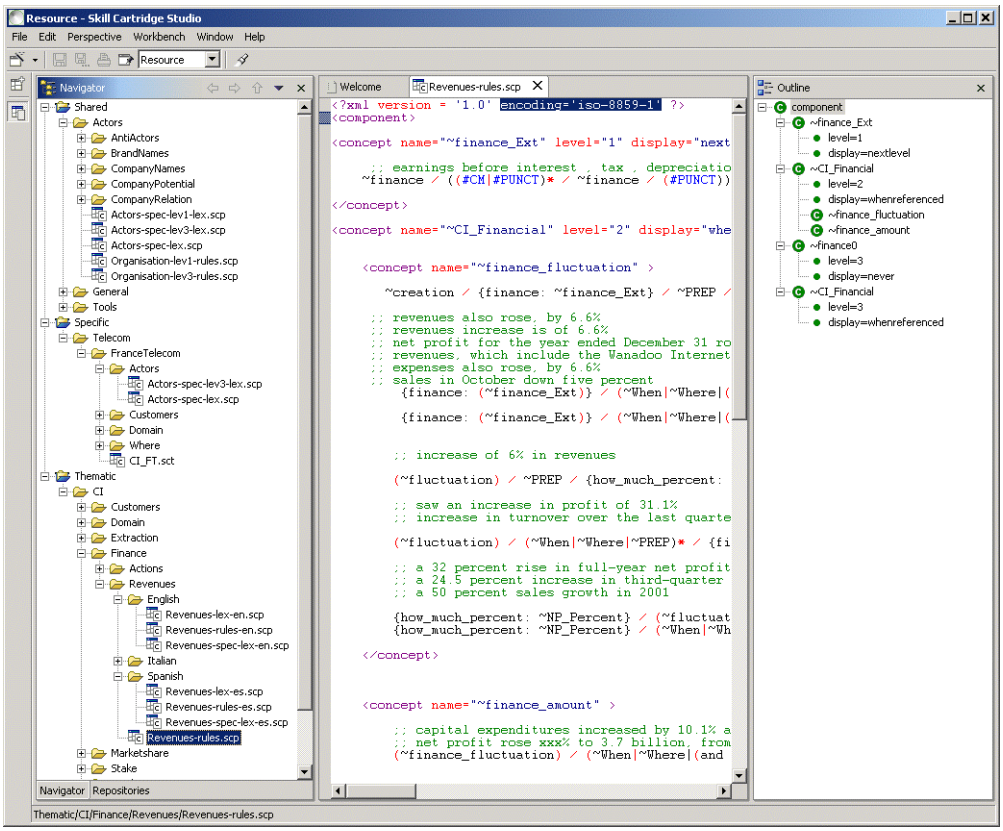

Figure 9 : la vue Edition