La normalisation : nouveau challenge en extraction d'information

Guillemin-Lanne Sylvie, Six Amandine

sylvie.guillemin–lanne@temis-group.com
amandine.six@temis-group.com

TEMIS
Tour Gamma B
193-197 rue de Bercy
75582 Paris cedex 12
FRANCE

Mots clefs :
Fouille de données textuelles, ingénierie des connaissances, extraction d'information, règle d'extraction, patron d'extraction, intelligence économique**,** veille scientifique et technologique, gestion des connaissances, ingénierie des connaissances, modélisation des connaissances, collecte d'informations

Keywords:
text mining, knowledge engineering, knowledge extraction, information extraction, extraction rules, extraction pattern., competitive intelligence, scientific and technical watch, knowledge management, knowledge engineering, knowledge modeling, information gathering

Palabras clave :
text mining, ingeniería del conocimiento, extraccion del conocimiento, extraccion de la informacion, reglas de extraccion, escudriñar científico y tecnológico, administración del conocimiento, ingeniería del conocimiento, formalización del conocimiento, reunir de información

Résumé

Cet article présente l'une des technologies de text mining développées au sein de la société TEMIS, à savoir l'extraction d'information. Il met l'accent sur l'évolution des objectifs auxquels doivent répondre nos solutions ceci grâce à la maturation conjointe des technologies TEMIS et des attentes des clients. L'extraction d'information remplit de plus en plus de fonctions dans le processus de traitement de l'information. Cette évolution soulève un nouveau challenge technologique: la normalisation. Notre propos sera illustré d'exemples concrets issus de deux applications réelles, une application de veille économique et concurrentielle, et une application de consolidation des textes de loi.

Abstract

This article presents one of the text mining technologies developed within TEMIS, namely the extraction of information. It stresses the evolution of the objectives which our solutions must answer in relation to the joint development of TEMIS technologies and customer expectations. Information extraction fulfils more and more of the needed functions in information treatment. This evolution raises a new technological challenge: standardization. Our presentation will be use concrete examples resulting from two real cases: an application using past and current economics texts and an application of the consolidation of law texts

Introduction

Notre article est une illustration des technologies de text mining développées au sein de la société TEMIS, et plus précisément celles d'extraction d'information. Contraction de Text Mining Solutions, TEMIS est un éditeur de logiciel qui conçoit et propose des applications dédiées à l'analyse textuelle, et liées à l'intelligence économique, à la gestion de la relation clients, à la gestion de la connaissance et des savoir-faire et à la gestion des ressources humaines.

La maturité conjointe de nos technologies et de l'expression des attentes des clients augmentent le niveau de complexité des applications mises en œuvre. Dans un processus de traitement de l'information que l'on pourrait conceptualiser en n étapes, l'extraction d'information, jusqu'alors centrée sur le repérage de l'information dans les textes, permet d'automatiser des étapes ultérieures, c'est-à-dire de :

- Structurer l'information
- Stocker l'information
- Exécuter les actions décrites par l'information elle-même

Passer de l'étape 1 (repérage de l'information) aux étapes ultérieures décrites ci-dessus soulève un nouvel enjeu : la normalisation de l'information extraite.

Après un court rappel sur le fonctionnement des Skill Cartridge™ couplées au serveur d'extraction d'information développés par TEMIS, nous présenterons les solutions mises en place pour répondre au besoin de normalisation. Nous illustrerons notre propos par deux cas clients, une application de veille économique et concurrentielle, et une application de consolidation des textes de loi et montrerons comment le processus d'extraction d'information maîtrisé permet l'automatisation d'un plus grand nombre de tâches liées au traitement de l'information.

1 L'extraction d'information

1.1 L'approche TEMIS

1.1.1 Skill Cartridge™

TEMIS dispose d'un serveur d'extraction d'information Insight Extractor™, couplé à des Skill Cartridge™ thématiques (intelligence économique), ou spécialisées par domaine d'activité (industrie pharmaceutique).L'ensemble fournit des applications dédiées à la veille stratégique et concurrentielle, à la gestion de la relation clients, à la gestion de la connaissance et des savoir-faire et à la gestion des ressources humaines.

L'information à extraire est modélisée et organisée selon une hiérarchie de composants de connaissance modulaires intégrables à différents domaines d'activité et/ou langues, appelée Skill Cartridge™ ou cartouche de connaissance. Un composant de connaissance peut avoir la forme d'un ou de plusieurs dictionnaire(s) ou d'un ensemble de règles d'extraction. L'objectif est de construire des patrons d'extraction (extraction patterns) [Yangarber et Grishman, 1997] suivant une approche guidée par le but [Appelt, 1993] [Poibeau, 2002]. Un patron d'extraction décrit une structure syntaxique de surface comportant des éléments lexicaux et/ou amorces (trigger words), des tags grammaticaux et des éléments typés sémantiquement.

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. Ces règles d'extraction combinent l'accès aux formes de surface, aux tags grammaticaux et aux lemmes, et aux concepts précédemment construits. En associant un concept à un patron d'extraction, 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. Le module d'extraction utilise la technologie des transducteurs [Hobbs 1997].

Notre approche de développement des composants de connaissance est guidée par l'idée de permettre l'exploitation de l'information extraite. Pour illustrer la méthodologie employée, nous nous concentrerons sur :

- L'arbre de concepts de la Skill Cartridge™
- L'architecture en plug-in.

1.2 La normalisation

1.2.1 Enjeux et périmètre

Jusqu'à présent les solutions développées par TEMIS avaient pour objet de repérer l'information pertinente dans les textes, puis de la mettre en valeur afin d'en accélérer sa lecture (information surlignée, ou présentée sous forme de synthèse dans un rapport html). Après une lecture rapide, les analystes pouvaient alors valider les informations extraites, celles-ci venant alimenter des fiches clients. Pour une illustration de telles applications qui répondent particulièrement aux problématiques de veille, nous renvoyons notre lecteur à notre article précédent [Delecroix et al. 2004].

Aujourd'hui, suivant les besoins exprimés par nos clients, notre tâche consiste à automatiser les étapes qui suivent le repérage de l'information, en exploitant le contenu de cette information.

Afin de traiter l'information extraite, de l'exploiter automatiquement pour, par exemple, instancier les champs d'une base de données ou la transformer en « ordre d'actions » à exécuter, il est nécessaire de structurer cette information. L'objectif est alors de conceptualiser au maximum l'information extraite en s'affranchissant de la façon dont elle est exprimée dans la phrase.

1.2.2 Solutions mises en œuvre

1.2.2.1 Arbre de concepts de la skill cartridge™

Pour répondre à l'objectif fixé, la première tâche consiste à normaliser les résultats d'extraction. Il s'agit ici de s'abstraire de la forme de surface pour restituer avec régularité une même information, un même type d'information sous un même format. Les résultats d'extraction s'exprimant sous forme d'arbres de concepts, l'enjeu est de produire la même structure d'arbre de concepts.

L'arbre de concepts est une structure XML restituant les concepts et leur contenu ainsi que les informations relatives à la position dans la phrase du texte encapsulé au sein du concept. Les noms des concepts, ainsi que leur hiérarchie sont définis par les règles d'extraction de la Skill Cartridge™. En intervenant sur les options d'affichage des concepts, il est possible, suivant les cas, d'alléger ou au contraire de détailler une hiérarchie de concept.

Les arbres de concepts offrent une certaine conceptualisation de l'information, l'objectif étant de produire des arbres qui, par leur hiérarchie, diffèrent aussi peu que possible. Or, la structure des résultats est directement liée à la structure des patrons d'extraction. Elle est en partie contrainte par la façon dont l'information est exprimée dans le texte, et en partie par la façon dont sont écrites les règles d'extraction. Les nouveaux enjeux d'automatisation nous conduisent naturellement à une réflexion plus poussée sur les choix d'affichage des concepts et sur l'architecture de la Skill Cartridge™ pour mieux maîtriser les arbres en sortie.

On doit tendre à produire des arbres de concepts identiques, même si la structure de surface est différente :

1) X a racheté Y pour 10 millions d'euros

/ Extraction Acquisition

Who X

/buy
Whom Y
Howmuch 10 millions d'euros

2) Z a été racheté par T en 2003

/Extraction Acquisition
Whom Z
/buy
Who T
When 2003

Le principe d'extraction d'information par modélisation des règles d'extraction implique que ces 2 exemples, aux éléments optionnels près, produisent le même arbre de concepts : *A rachète B*, les autres informations (Montant, Date) étant optionnelles. L'ordre d'apparition des concepts n'importe pas, seule la hiérarchie est importante. Il reste des cas, où il est difficile de produire un arbre identique, par exemple, lorsque la structure d'un concept est éclatée :

3) X augmente son CA de 20%

/Extraction Financial
who X
/CI Financial/augmente son CA de 20%

4) L'augmentation du CA de X de 20%...

/Extraction Financial
/Finance/augmentation du CA
who X
/Percent/20%

Pour être interprétable, l'arbre de concepts produit en *4* devra être réorganisé. C'est le rôle des plug-in de manipuler ainsi les arbres de concepts à l'issue de l'extraction d'information. Pour être aisément interprétable par des automates, les arbres de concepts devront être aussi simples et réguliers que possible, ce qui nécessite que 1) tous les cas soient couverts et que 2) les arbres de concepts varient le moins possible.

1.2.2.2 Architecture en plug-in

L'objectif des plug-in est d'interpréter les arbres de concepts et de les transformer suivant un modèle unique. Il s'agit de tâches java, activées à l'issue du processus d'extraction d'information, qui manipulent en entrée un arbre de concepts et renvoient un arbre de concepts modifié.

Ce qui est vrai à l'échelle de l'arbre de concepts, l'est également pour les concepts eux-mêmes. En effet, pour rentrer dans les champs d'une base de données, les informations telles que dates, montants financiers, etc. doivent être normalisées. Le plug-in répond à ce double objectif de normaliser les éléments de l'arbre et de simplifier l'arbre lui-même. On a notamment recours au plug-in pour procéder aux opérations de normalisation ou de développement.

Nous avons choisi, pour illustrer notre propos, un exemple standard : les dates. Il nous a semblé intéressant, dans la mesure où il permet d'exposer les actions successives effectuées par le plug in, à savoir :

- la normalisation,
- le développement.

La normalisation

Pour prendre place dans une base de données et être interprétables, les dates doivent avoir un format commun et de préférence chiffré, et ce quelle que soit leur forme d'expression dans les textes :

- 1er janvier 1983
- 01-01-1983
- 1 janv. 1983
- premier janvier 1983
- 01/01/1983

A l'issue de la phase d'extraction, la Skill Cartridge™ produit le résultat suivant :

5) 1er janvier 1983

/Date

/DAY/01	1er
/MONTH/01	janvier
/YEAR	1983

6) 01-01-1983

/Date

/DAY	1
/MONTH/01	janvier
/YEAR	1983

7) 1 janv. 1983

/Date

/DAY	1
/MONTH/01	janv.
/YEAR	1983

8) premier janvier 1983

/Date

/DAY/01	premier
/MONTH/01	janv.
/YEAR	1983

9) 01/01/1983

/Date

/DAY	01
/MONTH	01
/YEAR	1983

Dans cet exemple, on a veillé à découper le jour, le mois, l'année pour pouvoir les manipuler et les mettre au format choisi. L'organisation hiérarchique du lexique permet de passer du nom du mois à son numéro, et également d'interpréter l'adjectif numéral « premier ».

L'action du plug-in consiste à réordonner les concepts DAY MONTH et YEAR. Les différentes expressions de cette même date renvoient, après normalisation par le plug-in, au même arbre simplifié, les sous-concepts DAY, MONTH et YEAR, n'ayant plus lieu d'être, ont été supprimés :

/Date

01/01/1983

Le développement

Certaines dates nécessitent d'être développées avant d'être normalisées :

10) 13 et 19 septembre 2003

11) 24, 26 juin et 2 août 2004

L'arbre de concepts devant permettre de relier 13 comme 19 à septembre, et 24 comme 26 à juin.

Le plug-in consiste d'abord à développer et compléter l'arbre de concepts pour recomposer les dates au format DAY MONTH et YEAR, afin qu'elles puissent ensuite être normalisées.

Arbres de concepts en sortie d'extraction d'information :

10) 13 et 19 septembre 2003

/Date

/DAY	13
/DAY	19
/MONTH/09	septembre
/YEAR	2003

11) 24, 26 juin et 2 août 2004

/Date

/DAY	24
/DAY	26
/MONTH/06	juin
/DAY	2
/MONTH/08	aout
/YEAR	2004

Arbres de concepts après développement :

10) 13 et 19 septembre 2003

/Date

/DAY	13
/MONTH/09	septembre
/YEAR	2003

/Date

/DAY	19
/MONTH/09	septembre
/YEAR	2003

11) 24, 26 juin et 2 août 2004

/Date

/DAY	24
/MONTH/06	juin
/YEAR	2004

/Date

/DAY	26
/MONTH/06	juin
/YEAR	2004

/Date

/DAY	2
/MONTH/08	août
/YEAR	2004

Arbres de concepts après normalisation :

10) 13 et 19 septembre 2003

/Date

/13/09/2003

/Date

/19/09/2003

11) *24, 26 juin et 2 août 2004*

/Date
/24/06/2004

/Date
/26/06/2004

/Date
/02/08/2004

Ainsi les nouveaux enjeux d'automatisation nous ont conduit, non seulement, à une réflexion plus poussée sur les choix d'affichage et sur l'architecture de la Skill Cartridge™ à définir une méthodologie simple et efficace pour produire en sortie des arbres facilement interprétables. La phase de normalisation de l'arbre de concepts et des différents éléments qui le composent entre dans cette méthodologie.

Cette méthodologie nous a permis de répondre aux problématiques de nos clients dont nous choisissons de développer deux exemples d'applications réelles, l'une liée à l'intelligence économique, l'autre relevant du domaine juridique.

2 Une application de veille économique et ses évolutions

Ce premier cas client concerne un grand groupe pétrolier et gazier international, dont la Direction en charge de l'information pour l'ensemble du groupe a souhaité mettre en place une solution évolutive pour surveiller la concurrence et les marchés Pétrole / Gaz.

Il s'agit d'une application qui analyse les flux de presse en ligne, (Newswire), lit les flux quotidiens et propose un rapport journalier des informations extraites portant sur les capacités mondiales de production de pétrole. L'objectif étant de :

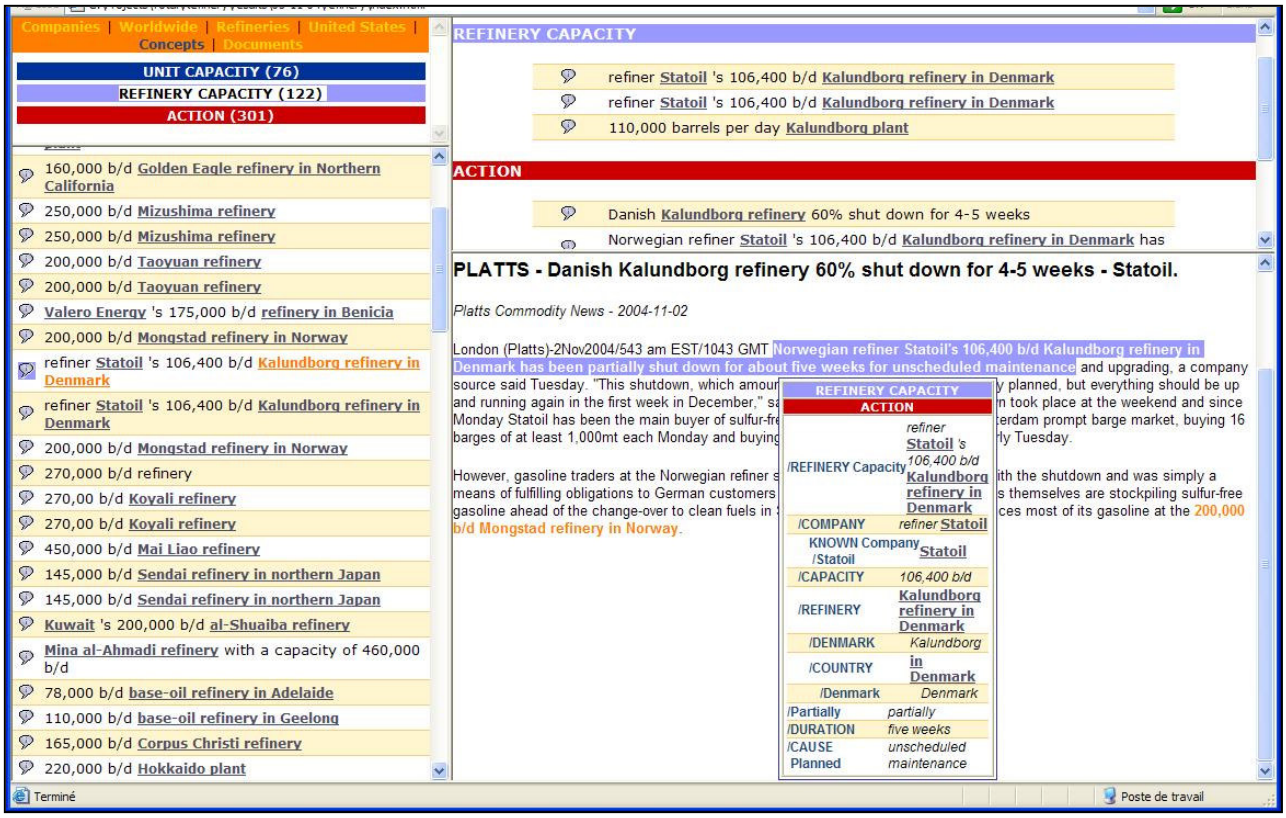
- Détecter les capacités mondiales de raffinage
- Générer automatiquement des rapports d'analyse quotidiens
- Offrir des outils de découverte d'informations stratégiques

Dans le cadre de ce projet, une première application a été déployée début 2004. Puis des applications spécifiques ont été développées en complément. TEMIS a notamment développé une Skill Cartridge™ dédiée pour

- détecter, normaliser et lier des sociétés et des raffineries
- détecter les capacités des raffineries et unités de production
- Détecter les arrêts et leurs causes : accident (feu), maintenance, événement politique

Une chaîne applicative a été instanciée pour:

- Télécharger les articles depuis les flux de presse en ligne (Factiva)
- Annoter les articles à l'aide de la Skill Cartridge™
- Stocker les articles annotés dans un historique
- Publier un rapport journalier des articles publiés la semaine précédente



Au bout d'un an d'utilisation de nouveaux besoins d'analyse et d'exploration des résultats sont apparus qui nécessitent la mise en œuvre d'une base de données permettant de recourir à des analyses temporelles ou croisées. Dans ce cadre, les résultats des Skill Cartridge™ ne mettent plus uniquement en relief l'information extraite mais alimentent dynamiquement une base de données. Les données telles que dates, montants financiers ou encore noms d'entreprises ont ainsi été normalisés.

3 Une application de consolidation des textes de loi

Ce deuxième cas client concerne un éditeur pour lequel TEMIS a développé en partenariat avec un intégrateur une application de consolidation des textes de loi à partir des arrêts paraissant au Journal Officiel. En effet, si les décisions de modifications des textes législatifs sont publiées au JO, les textes de loi eux mêmes ne font pas l'objet d'une republication avec leur mise à jour.

La consolidation des textes de loi est un enjeu majeur pour les éditeurs du domaine juridique. Cette opération jusqu'ici réalisée manuellement est longue et coûteuse. Il s'agit de :

- Lire le JO pour repérer les articles modificateurs
- Ouvrir le texte législatif qui doit être modifié
- Opérer la modification

Par ailleurs, les contraintes de nombre de caractères obligent les éditeurs de contenu juridique à être synthétiques. Ainsi souvent plusieurs articles sont cités, ce qui doit donner lieu à plusieurs renvois qui pointent vers des textes distincts :

11) *Article 12 et 13 du Code civil.*
→ Article 12 du Code civil
→ Article 13 du Code civil

Les renvois peuvent être factorisés sur les dates :

12) *Décrets des 21 janvier et 8 février 1991.*

ou sur les numéros

13) Lois n° 75-409 et n° 76-616 du 9 juillet 1976

Tous ces cas doivent être développés afin de créer autant de renvois complets que nécessaire.

La solution mise en place par TEMIS est intégrée à une plate-forme qui couvre d'autres fonctionnalités, notamment des fonctionnalités d'édition. TEMIS a développé une Skill Cartridge™ dédiée qui identifie :

- les renvois aux textes législatifs :
- Codes et Lois
- Articles, paragraphes, chapitres, sections...

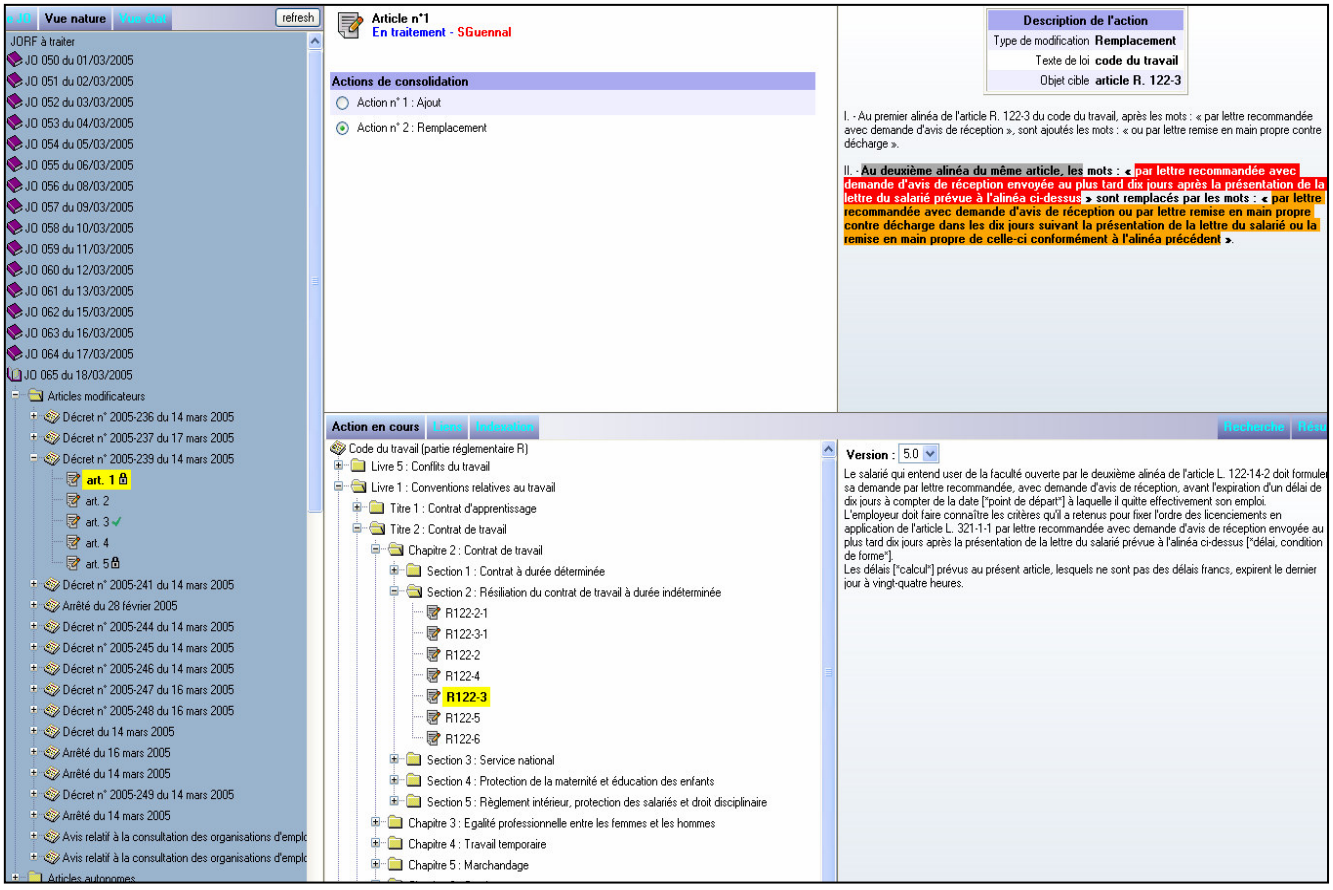
14) Articles 200 quater et 200 quater A du Code Général des impôts

- Les modifications
  - type d'action à opérer (suppression, ajout, remplacement)
  - contexte de la modification

15) A l'article 5 du Code Pénal, après les mots « … » sont ajoutés les mots suivants « … »

Les résultats de la Skill Cartridge™ sont structurés de manière à permettre l'automatisation de la modification. Le texte, l'endroit précis dans le texte et le type d'action à opérer sont clairement identifiés. Ces sorties d'extraction sont converties en tags XML, puis interprétés par l'application de consolidation qui procède alors aux modifications :

- Repérage de l'article à modifier dans le texte législatif concerné
- Repérage dans l'article des éléments à modifier (mots, phrases, paragraphes…)
- Application automatique de la modification
- Soumission de la modification aux éditeurs pour validation



4 Conclusion

Les technologies d'extraction d'information mises en œuvre chez TEMIS permettent de prendre en compte les profils clients et la diversité des objectifs de l'extraction d'information, et de développer des solutions adaptées à ces besoins. En se fondant sur une méthodologie identique, TEMIS a développé pour deux organismes totalement différents, deux solutions diversifiées, mais répondant parfaitement aux besoins exprimés par chacun d'eux.

Cet article met l'accent sur les nouveaux besoins adressés par l'extraction d'information, les problématiques de repérage de l'information sont désormais une étape nécessaire et non suffisante. Les tâches d'extraction doivent à présent être réalisées par des systèmes plus robustes qui structurent et normalisent les données. De ce point de vue, l'extraction d'information permet l'automatisation de tâches de plus en plus avancées dans le processus de traitement de l'information.

La stratégie de TEMIS s'inscrit dans le développement de composants de text mining modulaires pouvant facilement s'intégrer dans toute application client visant à extraire l'information pertinente à partir d'une collection de documents. Le nouvel enjeu défini est une porte ouverte à des applications plus intégrées toujours dans le but d'automatiser au maximum les tâches à faible valeur ajoutée pour concentrer les efforts humains sur les tâches d'expertise.

5 Bibliographie

[1] [Appelt *et al.* 1993] Appelt D., Hobbs J., Bear J., Israel D., Kameyama M. et Tyson M. « FASTUS : a finite-state processor for information extraction from real-world text ». I*n proceedings of the International Joint Conference on Artificial Intelligence* (IJCAI'93), Chambéry, 1993, pp. 1172-1178.

[2] [Aubry *et al*. 2002] Aubry Christophe, Grivel Luc, Guillemin-Lanne Sylvie, Lautier Christian « Aide à la construction de composants de connaissance pour l'extraction d'information : méthodologie et environnement» CIFT 2002 Colloque International sur la Fouille de Textes, Hammamet- Tunisie, 21-23 octobre 2002.

[3] [Buschbeck *et al*. 2002] Bushbeck Bianka, Grivel Luc, Guillemin-Lanne Sylvie, 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.

[4] [Delecroix *et al* .2004] Delecroix Bertrand, Guillemin-Lanne Sylvie, Six Amandine « Veille concurrentielle et veille stratégique : deux applications d'extraction d'information » VSST 2004 Veille Scientifique et Stratégique, Toulouse, 25-29 oct 2004,pp 117-128.

[5] [Eppstein 2001] Eppstein R. : Création d'un système d'information stratégique dans le domaine des technologies de l'information et de la communication – Application à CS Communication & Systèmes, Thèse, Université de Marne-la-Vallée, 28 novembre 2001.

[6] [Grishman 1997] Grishman R. Information Extraction: Techniques and Challenges. In M.T. PAZIENZA (éd.), *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, Springer Verlag, Heildelberg, 1997, pp. 10-27.

[7] [Grivel *et al.* 2001] 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 oct

[8] [Hobbs 1997] 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)

[9] [Neumann 1999] Neumann G., Schmeier S., Combining Shallow Text Processing and Machine Learning in Real World Applications. Proceedings of the IJCAI-99 workshop on Machine Learning for Information Filtering, Stockholm, Sweden, 1999.

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

[11] [Wilks 97] Wilks, Y.Information Extraction as a Core Language Technology. In Pazienza, M. T., editor, *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, Frascati, Italy, LNAI Tutorial, Springer. pp. 14-18,1997.

[12] [Yangarger et Grishman, 1997] Yangarber R., Grishman R., "Customisation of Information Extraction Systems". In Pazienza, M. T., editor, *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, Springer Verlag, Heildelberg, 1997, pp. 1-11.

[13] [Zanasi 2001] 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.