Intégration de Composants de Text Mining

Luc Grivel

INTEGRATION DE COMPOSANTS DE TEXT MINING POUR LE DEVELOPPEMENT D'UN SYSTEME DE RECHERCHE ET D'ANALYSE D'INFORMATION

Grivel Luc
TEMIS Text Mining Solutions
59, rue de Ponthieu 75 008 Paris
http://www.temis-group.com
luc.grivel@temis-group.com

Résumé : L'objectif de cet article est de montrer l'intérêt de l'emploi combiné de techniques d'analyse du texte (segmentation, lexicale, syntaxique, sémantique) et de diverses techniques d'accès à l'information (index, classification, catégorisation, cartographie) pour le développement d'un système de recherche et d'analyse d'information **qui soit adapté à des non-spécialistes des langages documentaires et qui s'intègre dans un processus de veille.** L'article montre comment ces techniques interviennent dans les fonctions d'un système d'analyse de l'information. L'originalité se situe dans l'approche (intégration de composants de text mining) qui est détaillée : reformatage XML des documents, visualisation des résultats, en passant par l'extraction des caractéristiques des documents et la classification.

Mots clefs : Fouille de données textuelles, extraction information, traitement du langage naturel, classification, hypertexte, cartographie

Abstract : The goal of this paper is to show the interest of combining various text analysis techniques (shallow parsing, semantic analysis, etc.) and some information access techniques (indexing, classification, clustering, mapping)) to developp an information analysis system to be used and customized by non-specialists of documentary languages. The paper shows how these techniques can be integrated to for a process chain including : XML reformating, information extraction, clustering, mapping.

Keywords : text mining, information extraction, natural language processing, classification, clustering, mapping, hypertext

Intégration de Composants de Text Mining pour le développement d'un système de recherche et d'analyse d'information

INTRODUCTION

L'objectif de cet article est de montrer l'intérêt de la combinaison de techniques d'analyse du texte (segmentation, lexicale, syntaxique, sémantique) et d'accès à l'information (index, classification, catégorisation, cartographie), rassemblées sous le nom de text mining, en prenant pour cadre le développement d'un système de recherche et d'analyse d'information **qui soit adapté à des non-spécialistes des langages documentaires et qui s'intègre dans un processus de veille**.

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

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

Comment les techniques citées plus haut interviennent elles dans ces fonctions?

Chacune des techniques est vue comme un composant aux fonctionnalités précises et délimitées. Chaque composant constitue un 'objet-serveur' ou objet distant et dispose d'une API (Application Programmatic Interface) Java publique lui permettant de s'intégrer facilement dans une application existante (Figure 1).

L'administrateur de sources 'crawle' différentes sources d'information, actionne des filtres de conversion XML des documents (en fonction du type de document et de la date de mise à jour) et les stocke dans un répertoire local.

Un serveur d'extraction pour dégager les concepts clé contenus dans les documents (noms de compagnies, dates, valeurs monétaires, fonctions, lieux, ou tout autre concept relatif à un domaine…) et génère les metadonnées décrivant chaque document.

Un serveur de recherche documentaire ou un SGBD stocke et indexe les documents et leurs metadonnées.

Un serveur incorporant un moteur de classification et un moteur de catégorisation classifie les documents (constitue des groupes), ou les catégorise (les place dans des groupes définis à priori).

DESIGN DU SYSTEME D'INFORMATION

Nous détaillons ici les composants et montrons comment ils interagissent dans cette architecture selon que l'objectif recherché est de naviguer dans une collection entière de documents ou de naviguer dans des résultats de recherche.

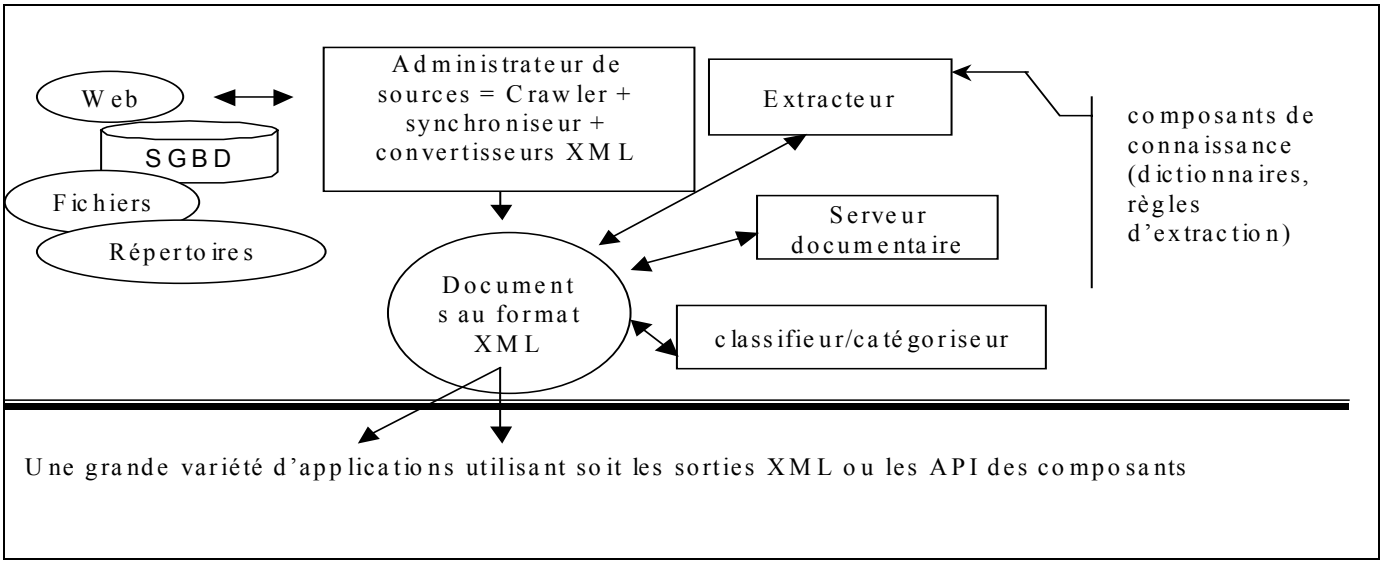

Figure 1 : Architecture

L'administrateur de sources

Il met à jour une base de données à partir de documents provenant de différentes sources pré-définies. Ce composant inclut les fonctions permettant de se connecter à différentes sources d'information, (Web, mail, news, data banks) dans différent formats (ascii, HTML, Word, PowerPoint, Excel, PDF) et de définir un profil de recherche sur ces sources. Tous les documents correspondant à ce profil sont périodiquement et automatiquement recherchés, copiés et stockés dans un répertoire local. Ils sont ensuite convertis dans un format homogène XML. Cette approche est aujourd'hui largement utilisée par les communautés des bases documentaires et des SGBD (orienté objet ou relationnel) lorsqu'il s'agit d'intégrer des documents hétérogènes. [MIC 98] [ABIT 97]

Dans l'exemple ci-dessous, une proposition d 'emploi comporte différents champs : Job Title, Description, Skills, Education, Location. Le reformatage XML a permis de conserver cette notion de zone de texte qui pourra être exploitée pour l'extraction et l'organisation de l'information.

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<doc_list xmlns:dc="http://purl.org/dc/elements/1.1/">
<DOC>
<dc:Identifier>bjo_1</dc:Identifier>
<dc:Title> OFFICE ADMINISTRATOR </dc:Title>
<text zone="location">
SC Missile Defense &amp;amp;Space Control - Anaheim
</text>
<text zone="description">
Site CM Office Administrator will be responsible for supporting one or more site resident CM work groups. Site CM OA performs various tasks contributing to office organization and efficiency. Duties included preparation and distribution of memos, reports, forms and business
 graphics; directs phone calls and electronic mail; schedules and coordinates meetings and business travel. Uses advanced computer SW features. Edits documents for correct grammar, punctuation, spelling, context and format.
   Anticipates what information and data will be needed for appointments, meetings and business travel. Follows up on assigned action items. Anticipates the need for office supplies. Establishes records, files and logs. This  position will include duties and assignments in support of the NMD program CM organization as required. This position is located in Arlington, VA.
</text>
<text zone="skills">
Team player with good oral and written communication skills. Nominally skilled in the use of the MS Office Suite of
    tools. Ability to work in a fast paced, multi-tasking environment.
</text>
<text zone="education">
Prefer vocational school training or equivalent work     experience; proficiency in operation of office equipment and business
software; and strong organizational, communication, and interpersonal skills.
</text>
</DOC>
…
```

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

Outre la récupération des meta-données existantes (par exemple, les champs qui sont présents dans le cas de documents provenant d'une base bibliographique), il est indispensable de générer des meta-données décrivant le contenu du document si on désire implanter des mode de recherche qui ne soient pas uniquement fondés sur une simple indexation de tous les mots du texte. C'est ici qu'intervient la phase d'extraction d'information.

Actuellement les systèmes d'extraction industriels les plus évolués s'appuient sur des analyseurs morpho-syntaxiques basés sur la technologie des transducteurs[1]. Ils prennent éventuellement en compte des règles d'extraction ou des ressources terminologiques. Les analyseurs morpho-syntaxiques sont de plus en plus performants (20 Mo/h pour Xelda de Xerox) et capables de traiter de plus en plus de langues différentes. Citons également les systèmes basés sur INTEX [Silb 99] et FASTR [Jacq 94] dans le monde universitaire.

Le serveur d'extraction développé par TEMIS est capable de traiter 7 langues (anglais, français, allemand, espagnol, hollandais, portugais, italien). Il est paramétré par un ou plusieurs composants de connaissances (Skill Cartridges™) qui définissent les éléments à extraire. Le système recherche les patterns décrits dans les règles et affecte aux éléments trouvés une étiquette syntaxique ou sémantique tels que des noms de compagnies, des relations (fusion de X avec Y, achat de W par Z), des noms de lieux, des dates, des prix, ….

Le résultat de l'extraction est une annotation du document (ici une offre d'emploi) par les éléments extraits. Le fait de connaître la sémantique associée à une zone de texte permet éventuellement de 'contrôler' les meta-données générées pendant la phase d'extraction d'information en développant des règles d'extraction adaptées à cette zone de texte. Par exemple, les fonctions dans un secteur d'activité donné, les diplômes requis, le nombre d'années d'expérience.

La méthodologie de construction de règles d'extraction développée par TEMIS est basée sur la notion d'organisation de règles d'extraction en niveaux hiérarchiques pour contrôler leur exécution sur les corpus[Bus 02]. Les problèmes relatifs à l'acquisition de la terminologie [Fau 00], [Bou 99], la création d'ontologies [Roc 00], la création de règles d'extraction [Gris97] sont complexes. Un programme de recherche [Gri 01a] 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 Cartridges™. Pour ce projet, TEMIS s'est vu décerner récemment le label technologie-clé par l'**ANVAR**.[Aub 02]

Organisation des données a priori versus organisation des résultats de recherche

A l'issue de l'étape d'extraction, chaque document est donc annoté par des concepts ou par des éléments syntaxiques (noms, verbes, adjectifs).

Les techniques usuelles pour faciliter l'accès à l'information sont :

- Recherche par mots-clés ou sur des concepts et classement des documents selon leur pertinence par rapport à la requête (Salton)
- Classification et cartographie: la classification (clustering) permet de regrouper les documents similaires par thèmes sans a priori sur la structure thématique. La Cartographie est un moyen de présenter un résumé de la classification,
- Catégorisation de documents : sur la base d'un vecteur de caractéristiques comprenant, entre autres, les résultats de l'extraction, la catégorisation permet d'affecter des documents à des rubriques ou catégories prédéfinies.
- Analyse statistique (distribution de concepts, distribution d'auteurs, distribution par dates de publication, …)

Ces techniques peuvent être combinées de manières différentes selon l'objectif recherché : naviguer dans une collection entière de documents ou naviguer dans des résultats de recherche.

Dans le cas où l'objectif est de permettre une navigation dans une collection de documents, il est nécessaire de présenter à l'utilisateur une structure fixe. Si on dispose pour chaque catégorie d'un ensemble de documents représentatifs, l'emploi d'un moteur de catégorisation[2] basé sur un modèle d'apprentissage s'impose-. Si l'on ne dispose pas d'exemples de documents pour chaque catégorie, le moteur de classification peut être utilisé. C'est l'approche qui était développée dans

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

[2] La Catégorisation ou classification supervisée peut être également utilisée pour prendre en compte un profil d'utilisateur. Ce dernier exprimant ses centres d'intérêt en fournissant un corpus de documents dont il a validé lui-même la catégorisation.

HENOCH (INIST) [GRIVEL 2000, 2001] et dans TEWAT (IBM) [COUPET 1995, 1998]. Après nettoyage et validation des résultats de classification sur la collection de documents, on peut utiliser ces mêmes résultats pour initialiser un moteur de catégorisation.

Dans le cas où l'objectif est d'aider à la navigation dans des résultats de recherche**,** la classification doit donc être effectuée à la volée, ce qui a un impact sur le choix de l'algorithme utilisé. Des exemples d'algorithmes adéquats sont donnés dans [Dou92]. Le serveur de classification de TEMIS est de type non hiérarchique (partition en K classes, K étant un paramètre fixé à priori), déterministe (on obtient le même résultat sur un même groupe de documents ordonnés). Il prend en entrée une description vectorielle des documents sous forme de hiérarchie de concepts et tient compte de cette hiérarchie pour la mesure de la similarité entre deux documents. Il est aussi possible de tenir compte des types de meta-données qui caractérisent les documents à classer, en excluant par exemple le contenu certains champs que l'on ne veut pas voir intervenir dans le calcul de similarité inter-documents.Une carte globale ou un tableau permettent de résumer les caractéristiques des clusters (les termes les plus pertinents du cluster, les relations inter-clusters, les statistiques sur différents champs des documents de chaque cluster (titres, auteurs, dates, …). Sur la base de ce résumé, l'utilisateur peut sélectionner un cluster qui lui semble contenir les documents les plus pertinents et classifier à nouveau ces documents s'il désire un niveau de détail plus fin sur la structure de ce sous-ensemble. Cette approche est particulièrement avantageuse lorsque l'utilisateur ne veut pas ou ne peut pas exprimer une requête bien formalisée, nécessitant de connaître toutes les finesses du langage de requête, la structure de la base documentaire, les plans de classement éventuellement utilisés (cas des brevets ou des offres d 'emploi). Il peut exprimer des requêtes plus simples, comportant éventuellement des termes ambigus et s'appuyer sur la classification pour sélectionner les documents les plus intéressants. Ceux-ci sont, le plus souvent, rassemblés au sein d'un même cluster [Hea96]. Les 'clusters-poubelle' sont rapidement identifiés (peu de documents, peu cohérents). Ce qui a pour conséquence de réduire le temps passé à l'identification des documents intéressants lorsque la réponse à une requête comporte trop de documents pour être analysée séquentiellement ou dans un temps limité.

L'exemple ci-dessous (Figure 2) montre comment sur un ensemble d'offres d'emploi aspirées sur un site Web, on peut à partir du résultat d'une recherche sur un terme général tel que 'management' distinguer les ensembles de document ayant trait au 'business management', au 'sofware management', 'risk management', 'project management', 'option and configuration management', …

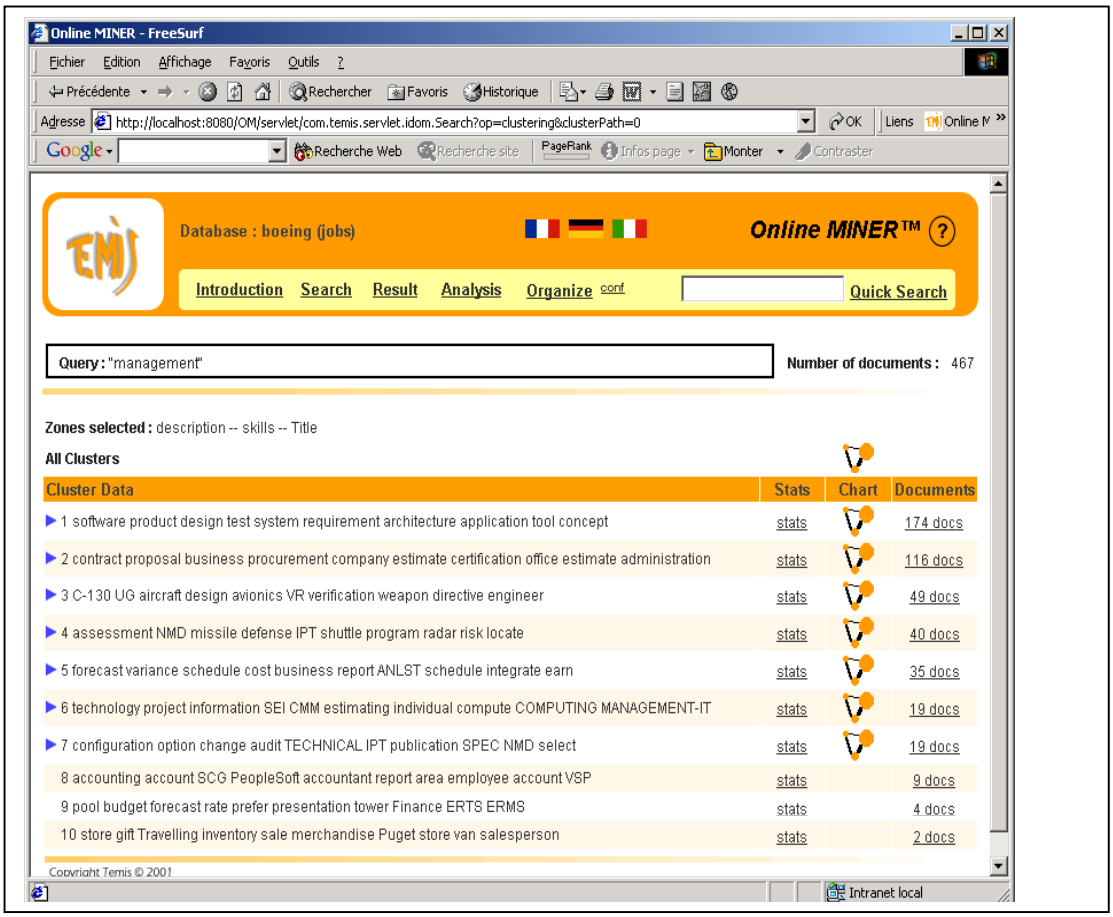

Figure 2 :Résultat d'une classification sur les champs 'description', 'skill' et 'titles'

CONCLUSION

Plus simples à développer, plus robustes car testés dans des contextes différents, ces composants peuvent être assemblés pour créer des applications dans des domaines variés et dans différentes langues. Citons par exemple :

- Grunher+Jahr en Allemagne, premier éditeur de presse en Europe, filiale de Bertelsmann : un système d'indexation et de catégorisation d'articles de presse,
- un consortium italien (TELCAL) : un système de veille dans le domaine de l'artisanat, l'agriculture et le tourisme,
- et des applications de gestion des relations humaines, de gestion de la relation clientèle ou de veille concurrentielle ou technologique en Suisse, Allemagne, USA, et France.

BIBLIOGRAPHIE

[ABI 97] ABITEBOUL S., CLUET S., CHRISTOPHIDES V., MILO T., MOERKOTTE G., SIMEON J. - Querying Documents in Object Databases -, International Journal on Digital Libraries, 1(1), 5-19, 1997.

[AUB 02] Aubry, C., Grivel, L., Guillemin-Lanne, S., Lautier, C., « Une méthodologie et un environnement d'aide à la construction de composants de connaissance pour l'Extraction d'Information » *CIFT'02, **Colloque International sur la Fouille de Texte 20-23 octobre 2002***, Hammamet-Tunisie, 2002.

[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.

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

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

[DOU 92] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey, *Scatter/Gather: ACluster-based Approach to Browsing Large Document Collections*, SIGIR '92, Pages 318 – 329, 1992.

[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.

[Gei 00.] Geißler S. "The DocCat system - Automatic Indexing in a pratical Application" in: "Medien Informations Management Praxis, Projekte Präsentationen" Verlag für Berlin-Brandenburg, Potsdam Hrgs. Ralph Schmidt Potsdam, 2000,Seiten 48-55

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

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

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

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

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

[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.

[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)

[HEA 96] Hearst M.and Pedersen J., Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results, *Proceedings of the 19th Annual International ACM/SIGIR Conference*, Zurich, August 1996

[MIC 98] MICHARD A. 'XML Langage et application' Editions Eyrolles, 361 p, 1998

[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.

[ROC 00] Roche C. « Corporate Ontologies and Concurrent Engineering », in: *Journal of Materials Processing Technology* volume 107, pages 187-193, Elsevier Science, 2000.

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

[SIL 99] Silberztein, 1999. *INTEX: a Finite State Transducer toolbox*, in Theoretical Computer Science #231:1, Elsevier Science

[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.

[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, (RFIA'2000), volume II, pages 101--110, Paris, France