Ontology Maintenance using Textual Analysis

Yassine GARGOURI (1), Bernard LEFEBVRE (2) and Jean-Guy MEUNIER (1)

(1) Laboratory of Cognitive Information Analysis
(2) Laboratory of Knowledge Management, Diffusion, and Acquisition
UQAM, Montreal (Quebec) H3C 3P8, Canada

ABSTRACT

Ontologies are continuously confronted to evolution problem. Due to the complexity of the changes to be made, a maintenance process, at least a semi-automatic one, is more and more necessary to facilitate this task and to ensure its reliability. In this paper, we propose a maintenance ontology model for a domain, whose originality is to be language independent and based on a sequence of text processing in order to extract highly related terms from corpus. Initially, we deploy the document classification technique using GRAMEXCO to generate classes of texts segments having a similar information type and identify their shared lexicon, agreed as highly related to a unique topic. This technique allows a first general and robust exploration of the corpus. Further, we apply the Latent Semantic Indexing method to extract from this shared lexicon, the most associated terms that has to be seriously considered by an expert to eventually confirm their relevance and thus updating the current ontology. Finally, we show how the complementarity between these two techniques, based on cognitive foundation, constitutes a powerful refinement process.

Keywords: Ontology maintenance, Latent Semantic Indexing, Singular Value Decomposition, classification, correlated terms, Vectorial Model and Cognitive psychology.

1. INTRODUCTION

Ontology is known in knowledge representation community as "a *formal and explicit specification of a shared conceptualization*" (Gruber, 1993). A domain ontology represents terms, specific to the domain, and their relations. It provides preliminary knowledge required for systematic information processing for navigating, recall, precision, etc. However, ontologies are continuously confronted to evolution problem. Due to the complexity of the changes to be made, a maintenance process, at least a semi-automatic one, is more and more necessary to facilitate this task and to ensure its reliability. Presently, there is still no consensus on methods and guidelines for such process. Nevertheless, texts are recognized as an essential source for ontology construction and domain knowledge.

How can we maintain a given explicit ontology in front of a dynamic world, characterized by continuously unstable textual data? How can we extract, from these texts, terms (or concepts) and their relations that are pertinent for an ontology and help maintain it? Because of the complexity of this problem, we will mainly deal in this paper, with only one dimension of this problem, which is the extraction of highly semantically related terms. Further dimensions, such the extraction of emergent terms in the texts that are related to certain ontologies, or the integration of new terms and relations with those of the current ontology, will be presented in our future work.

In this paper, we first put in context our technique by pointing out a brief approaches review. Then, we detail the different steps necessary for processing texts and extracting from them certain types of information pertinent for the maintenance of the ontology. Initially, we deploy the document classification technique to generate classes of texts segments having a similar information type and identify their shared lexicon, agreed as highly related to a unique topic. Further, we apply to these classification results the Latent Semantic Indexing technique associated with the Singular Value Decomposition, in order to extract from this shared lexicon, the most associated terms that has to be seriously considered for ontology maintenance. Finally, we discuss this approach and outline the future work related to the integration of the current ontology and the use of thesaurus.

2. APPROACHES REVIEW

Two classical methodologies are proposed for the semi-automatic analysis of large textual data to extract relevant knowledge, which are "*numerical methods*" and "*linguistic methods*". These two techniques are rather complementary. Because of its semiotic and linguistic characteristics, the traditional data-processing is usually linguistic. In fact, a text is seen as a succession of sentences that must be subject to linguistic analyzers. This approach seems completely natural since it corresponds in theory, to the human normal process of reading (Meunier, 1996). However, a real problem pertains to a theory of texts. Are texts linguistics phenomena? The answer depends on what one understands as "linguistics" is. If it is strictly understood as "grammatical", then a text is not a grammatical phenomenon. Although some authors think it is (Pavel, 1976, Dijk, 1977), others such as (Rastier and al., 1994; Meunier, 1996) content such a view.

It seems that numerical approaches then allow to extract much more regularities in text than the strict linguistic (grammatically based) ones. The numerical approach, especially based on classifying strategies, allows a considerable saving of time during corpus exploration, and for this reason, they are essential when confronted with vast textual corpora. In addition, they are extremely useful to quickly detect semantic and textual associations. Moreover, when associated with additional

resources such as thesauri[1], they deliver a precious assistance for global analyses.

With regard to the ontology maintenance precisely, several techniques for natural language processing, information extraction, machine learning and data mining are used to extract concepts from corpus. These techniques are relatively "mature". However, terms relations extraction is obviously a more complex and difficult problem, as shown in various projects such as "InfoSleuth" (Hwang, 1999), "Scalable Knowledge Composition" (Jannink & Wiederhold, 1999), "Ontology Learning" (Maedche and Staab, 2000), "Inductive Logic Programming" (Lavrac and Dzeroski, 1994), "University Michigan Digital Library" (Weinstein, 1998).

3. PROPOSED METHODOLOGY

The ontology construction refers mainly to the identification of relations between terms and concepts. As a consequence of the evolution of the ontology, we have to explore, in the texts, the related terms that seem pertinent for the current ontology modification. Then, theses terms and there relations have to be consistent with those of the current ontology.

Our objective in this paper consists of setting up a methodology that supports the user through the discovery of terms relations, that are potentially useful for the ontology maintenance. It's easy to put in issue automatic-based systems pretending to accomplish this task without any noise or imperfections. It seems more reasonable to follow a semi-automatic process involving a light expert domain intervention in some steps especially for result validation.

Texts constitute a tangible support, gathering stabilized knowledge which is used as reference, as well as a valuable knowledge resource. Moreover, the access to the terms and texts, justifying the definitions of the concepts ensure a better readability of the model and thus facilitate the ontology maintenance. Therefore, mining terms and their relations from texts is attracting increasing attention in the knowledge management community (Meunier and al., 1999).

Through a set of processing operations on texts, we aim to extract valuable relations between terms. In the first phase, a numerical method is used to quickly select groups of terms, that are potentially related, and that deserve more refinement processing in order to extract couples of strongly related terms. This task is accomplished in a second phase, using the technique of Latent Semantic Indexing approach (LSI) (Deerwester and al. 1990, Srivastava and al. 2002), which helps to identify, among each class of terms, the most correlated ones. All relations between these terms are to be considered by an expert to confirm their relevance in the updating of the current ontology.

For numerical classification purpose, we mainly use a neural net approach embedded in a software called **GRAMEXCO**, which is an instance of a sequence of modules built from a generic platform called **SATIM**[2] (Biskri, I., Meunier, J.G. 2002). As a computer system for textual information processing, this platform allows exploration and experimentation of various types of analysis due to the modularity, its many analysis functions and its sensitivity to the growth of raw textual data. In particular, **GRAMEXCO** allows executing a data processing sequence on texts to classify the segments, which is based on the N-grams approach (Damashek 1989).

We detail in the following sections our bootstrapping process which is organized in six major steps:

Step 1 : Tri-grams extraction and terms filtering

Using **GRAMEXCO**, the first step consists of extracting N-grams of characters from text and identifying segments (i.e. parts of document). Theses N-grams are defined as a sequence of N characters (for instance, sequences of three characters are called tri-grams). These two objects form the matrix that has to be used by the classifier. In other words, textual segments are compared and classified on the basis of N-grams co-occurrence.

Analyzing a text in terms of N-grams, constitutes a valuable approach for text written in any language based on an alphabet and the concatenation text-construction operator. Clearly, this is a significant advantage over the problematic notion of what a word is. In addition, the use of N-grams of characters instead of words offers another important advantage: it allows controlling the size of the lexicon used by the processor, as shown in (Lelu and Halleb, 1998).

The "*term extractor*" (a module of **GRAMEXCO**) is used to identify the lexicon (set of lexemes) from a corpus. Whereas N-grams simply serve to the classification purpose, the lexicon plays more effective role in the following steps. As a result, and before processing this lexicon and extracting N-grams, some important filtering operations have to be conducted to guarantee more reliable results. In other words, an automatic lemmatisation process is applied to replace terms by their lemma. In fact, terms such as {*inform, information, informing,…*}refer to the same concept; «*information*», and should therefore, be analysed as a unique term in the following steps. In addition, a filtering process makes it possible to eliminate functional terms (known as "*stop-words*" or "*trivial-words*") such as {*the, a, in, at…*}, as well as semantically insignificant terms. Indeed, very frequent terms and hapax don't play an effective role for segment discrimination (although they could be important for ontology maintenance as we will discuss later). Despite the fact that keeping these terms does not affect drastically the classification process, it could however generate some noise for the following processes. It's clear here, that important choices have to be done by the user. For this reason, **GRAMEXCO** offers flexibility and conviviability to support his tasks.

Step 2 : Classification and identification of related terms

The aim of the classification is to extract some type of "*semantical regularities*" between segments of the text (Manning,

[1] The work described here is in progress. In particular, experiments related to the use of thesaurus is not described here. Further details will be available in our future work.

2 SATIM can process other kind of information than texts, such as : images, sound.

and Schütze, 1999; Sebastiani, 2002; Gelbukh and al., 1999). These segments contain a similar information type and serve, as a consequent, to detect valuable indicators for associations between terms. As a textual data-mining method, this process, often less detailed than the linguistic and conceptual approaches, allows a first general and robust exploration of the corpus. It identifies the classes of segments and groups of lexemes having associations known as co-occurrence and thus, detects their semantic networks (Church and al. 1989, Lebart and Salem 1988, Salton 1988). It's typically processed using a numerical classifier as explored in (Meunier & al., 1997) (Memmi et al 1998) (Benhadid & al., 1998) (Biskri & Delisle, 1999). For **GRAMEXCO**, an ART (*Adaptive Resonance Theory*) (Grossberg, 1988) neural network has proven its effectiveness to computes similarities between segments and produce classes.

The classification uses as input a vectorial model that considers a text as a whole and aims to infer from texts, an implicit semantic structure (Salton & McGill 1983). This model translates a text into a space matrix that correlates texts segments with terms (here N-grams) and thus produce networks of terms corresponding to the topics treated in the text (Memmi, 2000). Indeed, the co-occurrence of terms within different parts of the text supports intensely the coverage of the same topic. The vectorial model generally follows typical steps: vectorization, classification, interpretation and use. The vectorization step, which is the text representation using vectors, needs a fundamental choice of the selected elements and the representative characteristics. This choice is obviously dependent to the subsequent tasks that have to be done. For our case, the pre-processing steps are crucial for detecting new terms associations that could be added to the current ontology.

Since the classification process leads to gathering terms related to a particular topic, it discovers the different contexts a term may possesses in text, hence identifying its precise semantical environment in which its meaning is deployed and allowing to resolve, to some extent, the lexical ambiguity problem.

After segments classification, we extract the lexicon from each class, which is taken to be the intersection of the terms belonging to all segments of the class. The co-occurrence of terms within some particular segment is considered as serious indicators for strong terms associations. In other words, inside each class, terms are assumed to be potentially related. The following steps consist in determining more precise associations between couples of such terms.

Step 3 : Calculating the Terms weights

At this level, we aim to extract from these classes of terms, those representing a high level of correlation. Latent Semantic Indexing approach (LSI) uses such a technique (Deerwester and al. 1990, Srivastava and al. 2002,). It has been applied since the Nineties for semantic information retrieval from texts, although the first works on co-occurrence have started from the Seventies. This technique was chosen basically for its simplicity, and its basis on a fairly precise mathematical foundation.

This LSI technique uses as input a term document matrix corresponding to the terms weights. We point out here that the corpus is assumed as a collection of documents. So, we calculate a weight for each term using its occurrence in the document. Its value is given by the following formula:

$$w_{i,k} = \frac{C_{i,k}}{\sum_{j=1}^{n_k} C_{j,k}}$$

Where $w_{i,k}$ is the weight of term T_i in document D_k, $C_{i,k}$ is the number of occurrences of term T_i in the document D_k, and n_k is the total number of terms in the document. These terms are limited to those filtered and selected for classification.

Step 4 : Building the term document matrix

The statistics for each individual document are combined in order to produce a statistical analysis of the whole collection. A standard normalization for document length, as explained in (Greengrass, 1997), is used to avoid the fact that a term may have a large weight simply because the document in which it occurs is small, rather than because of its high frequency throughout the document collection. The normalized term weights become :

$$W_{i,k} = \frac{w_{i,k}}{\sqrt{\sum_{j=1}^{n_k} w_{j,k}}}$$

For each class of terms c, the corresponding weights form the matrix W_c. This matrix is referred to as the term document matrix, with rows representing the collection of m documents (D_k), and columns representing n_c terms (T_i^c) belonging to the class c.

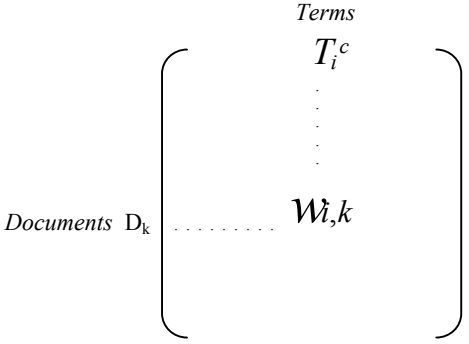

Figure 1 : The term document matrix W_c

Step 5 : Singular Value Decomposition

Inside each class of terms, we wish to determine couples of related ones. The LSI method links terms into a useful semantic structure, as described thoroughly by (Berry and al., 1995). LSI represents documents by concepts that are truly and statistically independent in a way that terms are not. A concept is considered here as a set of related terms. The LSI primarily involves

decomposing the matrix W_c using the Singular Value Decomposition (SVD) (Golub and al., 1969), which is a kind of linear regression. Thus, Wc can be decomposed as follows :

$$W_c = U \Sigma V^T$$

Where U is an ($m \times r$) term matrix, V is an ($r \times n_c$) document matrix, and Σ is an ($r \times r$) matrix, where r is the rank of W_c. Σ is a diagonal matrix containing the singular values of Wc. In this decomposition, the singular value σ_i corresponds to the vector u_i (the i^{th} column of U), and vi (the i^{th} row of V). The columns of U, the rows of V, and the diagonal values of Σ have been arranged so that the singular values are in descending order, moving down the diagonal. This formula transformation doesn't cause any lose of generality.

Step 6 : Extraction of potentially related terms

In this step, we consider extracting from each class of terms, the most related ones. Indeed, only these terms and their relations are to be considered for the ontology maintenance.

As discussed by (Deerwester and al. 1990; Nicholas and al. 1998), using LSI, we remove all singular values from Σ which fall below a threshold percentage of the largest singular value, σ_1. As a result, Wc can be approximated by W_c^s, with increasing accuracy as *s* approaches *r* :

$$W_c^s = U^s \Sigma^s V^{sT}$$

Where, Σ^s is derived from Σ by removing all but the largest *s* singular values, Us is derived from U by removing all but the *s* columns corresponding to the largest singular values, and V^s is derived from V by removing all but the s corresponding rows, where $s \leq r$.

U^s seems to be the most important component for us. Indeed, this ($m \times s$) matrix represents correlations between terms in the document collection and belonging to the class *c*. Each column of this matrix, u_i, is a vector, which we consider to represent a concept. The elements of u_i, give the correlation of terms to the concept. Zeroing out all elements in u_i, which fall below a certain threshold percentage of the highest correlated term in u_i, eliminates the weakly associated term (see Figure 2).

At the end of this process, all possible couples among the remaining terms, are assumed to be potentially related. The decision, whether or not to consider these new discovered relations between terms, has to be taken by a domain expert.

The steps 4 to 6 are repeated for each class of terms. Finally, a set of terms and their relations is delimited. However, their integration into the current ontology for maintenance purpose, is a relatively complex problem, that we will explore in our future work.

4. DISCUSSION

Researches on cognitive psychology show that most of words are assimilated by reading (Landauer and S.T. Dumais, 1997). Being exposed to texts, a learner tries, during his reading process, to refine gradually the word meaning using joint occurrences of these words with others. For example, in the absence of an explicit definition of the word "*microprocessor*", the learner is able, throughout his texts reading, to acquire the meaning of the word because this meaning is confirmed in the context in which this word appears with others such as "*card*", "*computer*", "*electronics", "hardware*", "*Central Processing Unit*", etc. However, a simply repeated co-occurrence of one word with others seems to be insufficient for its meaning acquirement. All the co-occurrences of all the words are rather required through the texts.

Based on this cognitive foundation, we support the idea of applying the classification technique to identify groups of terms appearing together and having semantical relations or, at least, semantical similarities when used in comparable contexts.

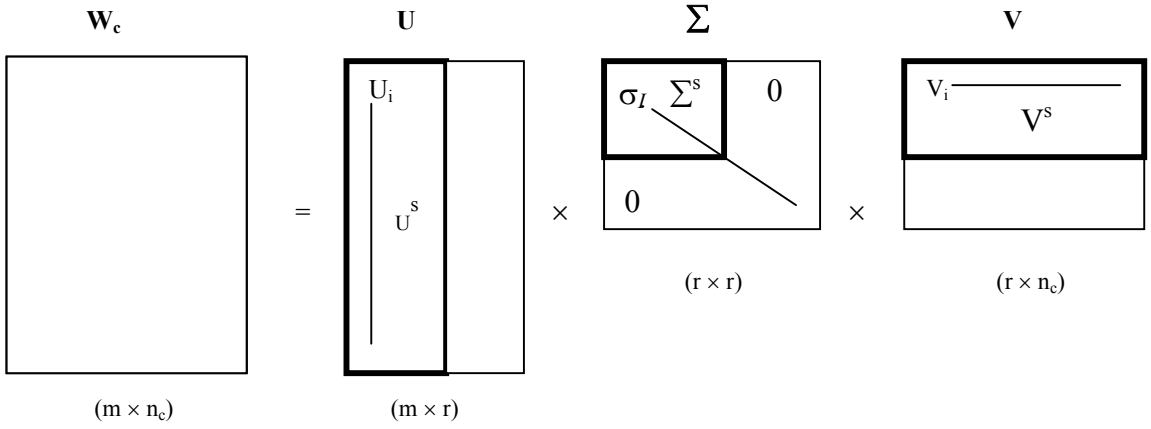

Figure 2 : Singular Value Decomposition of The term document matrix W_c

In addition, applying the LSI on the groups of terms identified by the classification, rather than on all the terms of the text, has the advantage to reduce the matrix of co-occurrence of terms in documents to a reasonable dimension. Indeed, in spite of the difficulty to identify in theory an adequate and precise dimension of this matrix, it's easy to prove that a huge dimension would prevent us from revealing sufficient semantical relations between terms, and also, a too small one would lead to a too great loss of information (Deerwester and al., 1990). In addition, this technique is automatic and domain independent. It's also applicable to large textual data.

Our approach has another important advantage due to its independence to a specific language. In fact, we privilege the N-grams technique for classification, as well as statistical basis for the LSI, which are both independent of the language, especially the syntactic characteristics. It's important to notice that nowadays, a model which is not multilingual would be controversial and as a consequence, of a restricted application.

However, with this approach, we don't pretend to extract the totality of related terms. Indeed, terms belonging to different classes may be related, but not considered in our case. For example, some frequent terms (or, on the contrary, hapax) don't appear among the classes of terms generated by **GRAMEXCO** and are consequently excluded from the following steps. Indeed, these terms have no effect on the discrimination of segments of texts. Whereas these terms have no significance in the classification process, their exclusion from the subsequent processes could be seen at first sight as controversial. However, this problem is not really "critical" since we consider the ontology maintenance as an iterative process, carried out on multiple text collections, in such a way that, terms appearing so frequently in some texts, do not in others (and the opposite for hapax) and they could, as a consequence, take place potentially in other maintenance processes.

Despite the success of the classification, this technique suffers from two serious limitations. First, the model can only handle stable corpora. In other words, if the texts change, all the process must be redone. Therefore, ART has been privileged in **GRAMEXCO** to deal efficiently with this dilemma. Second, the results produced by the classification are occasionally problematic in the absence of the linguistic interpretation. In fact, associations between terms belonging to the same class are not always clear or unambiguous. However, we think these limitations should not prevent researchers to consider classifiers for textual data mining purposes.

Compared to symbolic semantical representations such as semantical networks and terminological networks, the LSI approach has the disadvantage of generating only one kind of terms relations. As a consequence, an expert has to specify the semantical relations between terms judged as correlated by the LSI. Moreover, it's easy to imagine an interface where this technique proposes a terminological network that the expert gradually labels.

5. CONCLUSION AND FUTURE WORK

The proposed model constitutes a significant help in the field of ontology maintenance. It assists terminologists, charged to navigate through vast textual data in order to extract and normalize the terminology. In addition, it facilitates, the task of the knowledge engineers charged to model domains.

We show that our method is promising in its ability to extract reasonably good associations between terms. Indeed, the complementarity between, on one side, the document classification technique, in our case essentially based on neural networks (ART), and on the other side, the Latent Semantic Indexing approach, constitutes a powerful refinement process.

For the ontology maintenance, the complete and accurate identification of terms in a specific domain or corpus is considered as a pre-processing of the highest importance for the production of adequate and reliable results. As a consequence, specific techniques have to be considered to evaluate the reliance of these terms compared with others from the current ontology and also from additional sources of information such as thesauri.

Knowledge available in corpora is typically explicit and thus requires implicit knowledge. In Artificial Intelligence, knowledge has to be declared to support inductive processing. Thesauri are especially useful for offering lexical networks and additional information related to the term meaning (use, definition, synonymy, etc.). Therefore, we plan to expand on this work in the future by addressing the concerns raised in the use of this source of information. We also intend to expand our algorithm to possibly assist the user throughout his task of relations terms labelling using the thesaurus. Finally, our goal is to ensure a continuous ontology refinement, keeping in mind the consistency and the coherence of this ontology and its artefacts. This refinement process is under implementation as a new sequence of modules inside **SATIM**.

6. ACKNOWLEDGEMENTS

This work is being carried out under the KMDT Project (Knowledge Management and Diffusion in Telecommunications), by researchers from UQAM (University of Quebec at Montreal) and UofM (University of Montreal): Bernard Lefebvre, Jean-Guy Meunier, Gilles Gauthier, Omar Cherkaoui, Olivier Gerbé, and others master and PhD students.

We would like to thank LUB (Laboratoire Universitaire de Bell) and NSERC (Natural Sciences and Engineering Research Council of Canada) for their financial support and Ismail Biskri (from LANCI) for the platform SATIM (developed jointly with Jean-Guy Meunier).

7. REFERENCES

Benhadid, I., Meunier, J.G., Hamidi, S., Remaki, Z. and Nyongwa, M. (1998), "**Étude Expérimentale Comparative des Méthodes Statistiques pour la Classification des Données Textuelles**", Actes de JADT-98, Nice, France.

Berry, M. W., Dumais, S. T., O'Brien, G. W. (December, 1995) "**Using linear algebra for intelligent information retrieval**". SIAM Review. 37(4), 573-595.

Biskri, I., Delisle, S. (1999) "**Un modèle hybride pour le textual data mining : un mariage de raison entre le numérique et le linguistique**" . TALN 99, France, pages 55-64.

Biskri, I. and Meunier, J.G. (2002) "**SATIM : Système d'Analyse et de Traitement de l'Information Multidimensionnelle**", Proceedings of JADT 2002, St-Malo, France, 185-196.

Church, K., Gale, W., Hanks, P., Hindle, D., (1989). "**Word Associations and Typical Predicate-Argument Relations**", Proceedings of the 1st International Workhop on Parsing technologies, Carnegie Mellon University.

Damashek, M. (1989) "**Gauging similarity with n-grams: language independent categorization of text**". Science 267.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990) "**Indexing by latent semantic analysis**". JASIS, 41(6), 391-407.

Dijk, T.V. **(**1977) "**Text and context**". London: Longman.

Gelbukh, A., Sidorov, G. and Guzmán-Arenas, A. (1999) "**Text categorization using a hierarchical topic dictionary**". Proc. Text Mining workshop at 16th International Joint Conference on Artificial Intelligence (IJCAI'99), Stockholm, Sweden, July 31 – August 6, 1999, pp. 34-35.

Golub, G. H., Reinsch, C. (1969) "**Singular value decomposition and least squares solutions**". Handbook for Automatic Computation, Springer-Varlag, New York, 134-151.

Greengrass, E. (1997) "**Information retrieval: an overview**". NSA reports R521, 18-41.

Grossberg, S. (1988) "**Neural Network and Natural Intelligence**". Cambridge: MIT Press, 1988.

Gruber, T.R. (1993) "**A translation approach to portable ontology specifications**". Knowledge Acquisition, 5, 199-220.

Hwang, C. H. (1999) "**Incompletely and imprecisely speaking: Using dynamic ontologies for representing and retrieving information**". In. Proceedings of the 6th International Workshop on Knowledge Representation meets Databases (KRDB'99), Linköping, Sweden, July 29-30, 1999.

Jannink, J. and Wiederhold, G. (1999) "**Ontology Maintenance with an Algebraic Methodology: a Case Study**", in Proceedings of 1999 AAAI workshop on Ontology Management, Orlando FL.

Landauer T.K. and Dumais S.T. (1997) "**A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge**". Psychological Review vol 104(2), pages 211-240, 1997.

Lavrac, N. and Dzeroski, S. (Eds.) (1994). "**Inductive Logic Programming: Techniques and Applications**". Ellis Horwood.

Lebart, L., Salem, A. (1988), "**Analyse statistique des données textuelles**", Paris: Dunod.

Lelu A., M. Halleb & B. Delprat (1998). "**Recherche d'information et cartographie dans des corpus textuels à partir des fréquences de n-grams**", Proceedings of JADT-98, Nice, France.

Maedche, A. and Staab, S. (2000) "**Discovering conceptual relations from text**". In, ECAI 2000. Proceedings of the 14th European Conference on Artificial Intelligence, IOS Press, Amsterdam.

Manning, C.D., Schütze, H., (1999), "**Foundations of Statistical Natural Language Processing**", MIT Press.

Memmi, D., Meunier, JG. and Gabi, K. (1998) "**Dynamical Knowledge extraction from texts by Art Networks**". Proceedings of Neurap. Marseille. 1998. p. 205-210.

Memmi, D. (2000) "**Le modèle vectoriel pour le traitement de documents"**, *Cahiers Leibniz n° 2000-14*.

Meunier, J. G. (1996) "**La théorie cognitive: son impact sur le traitement de l'information textuelle**". In V. Rialle et Fisette, D (Ed.), Penser l'Esprit, Des sciences de la cognition à une philosophie cognitive. (pp. 289-305). Grenoble: Presses de L'Université de Grenoble.

Meunier, J.G., Biskri, I., Nault, G., Nyongwa, M. (1997), "**Aladin et le Traitement Connexionniste de l'Analyse Terminologique**", Actes de RIAO-97, Montréal, Canada, 661-664.

Meunier, J.-G., Remaki, L. Forest, D. (1999) "**Use of classifiers in computer-assisted reading and analysis of text (CARAT)**". Actes du colloque international CISST 1999, Las Vegas, Nevada, U.S.A, p. 437 – 443.

Nicholas, C., Dahlberg, R. (1998) "**Spotting topics with the singular value decomposition**". Principles of Digital Document Processing, FIARA, St. Malo.

Pavel T., (1976) "**Possible worlds in Literary Semantics**" The Journal of Esthetics and Art Criticism*,* 34: 2.

Rastier, F., Cavazza, M., and Abeillé, A. (1994). "**Sémantique pour l'analyse, de la linguistique à l'informatique**". Paris,: Masson.

Salton G. and McGill M. (1983) "**Introduction to Modern Information Retrieval"**, McGraw-Hill.

Salton, G. (1988), "On the Use of Spreading Activation", Communications of the ACM, Vol 31 (2).

Sebastiani, F. (2002) "**Machine learning in automated text categorization**". ACM Computing Surveys, 34(1):1-47.

Srivastava S., Gil De Ladadrid, J. and Elvadapu C.S. (2002) "**Document Ontology: A Statistical Approach**". SSGRR'2002, L'Aquila, Italy.

Weinstein, P. (1998) "**Ontology-based metadata: transforms the MARC legacy**". In. Akscyn, F. & Shipman, F.M. (Edit). *Digital Libraries 98, Third ACM Conference on Digital Libraries*. New York: ACM Press, 254-263.