6. DEXA Workshop 1995: London, United Kingdom

Norman Revell, A. Min Tjoa: 6th Int. Conf. and Workshop on Database and Expert Systems Applications (DEXA'95) - Workshop Proceedings. ONMIPRESS, San Mateo, California

Object-Oriented Databases

- Thomas A. Mück, Martin L. Polaschek: A New Approach to Associative Access in OODB. 1-10
- Xuequn Wu, Jan Neuhaus, Q. D. Huynh: A Simple Yet Powerful Application Programming Interface for an Object-Oriented Database System. 11-18

Expert and Knowledge Based Systems I

- José Hilario Canós Cerdá, María del Carmen Penadés, Isidro Ramos, Oscar Pastor: A Knowledge-Base Architecture for Object Societies. 19-24
- Magdi N. Kamel, M. J. McCaffrey, P. G. Metzler: Design and Implementation of a Prototype Maintenance Advisor Expert System for the MK92 Fire Control System. 25-34

Active and Temporal Aspects I

- José Palazzo M. de Oliveira, Nina Edelweiss, Eduardo Arruda, Alberto H. F. Laender, João M. B. Cavalcanti: Implementation of an Object-Oriented Temporal Model. 35-44
- S. Castangia, Giovanna Guerrini, Danilo Montesi, G. Rodriguez: Design and Implementation for the Active Rule Language of Chimera. 45-54
- Cristina De Castro: Temporal Aspects in Distributed Relational Databases. 55-62

CIM and AI

- John E. Gallety, R. C. Daniel, A. Ikonomou: Recent Heuristic Methods for Production Scheduling. 63-76
- Heimo H. Adelsberger, Klaus-Peter Keilmann: Constraint Handling in Planning Systems for Manufacturing. 77-86

Object-Oriented Databases II

- Mohsen Beheshti, André de Korvin: Logical Optimization when Uncertainty is Present. 87-96
- E. Hong: A Schema Evolution Mechanism. 97-104

Expert and Knowledge Based Systems II

- Elizabeth A. Kemp, Elisabeth G. Todd, Damian Pacitto, David I. Gray: Developing an Expert System to Assist New Zealand Dairy Farmers. 105-114
- Mara Nikolaidou, D. Lelis, Dimosthenis Anagnostopoulos: Distributed System Design: An Expert System Approach. 115-123

Active and Temporal Aspects II

- Vangalur S. Alagar, Fereidoon Sadri, Joseph N. Said: An Extended Relational Model for Managing Uncertain Information. 257-266
- J. H. ter Bekke: Meta Modeling for End User Computing. 267-273
- Martin Gogolla: Towards Schema Queries for Semantic Data Models. 274-283

Legal Systems I

- Antonio Cammelli: Models for the Logic Representation and Automatic Interpretation of a Legal Text. 284-290
- Rosa M. Di Giorgi, Elio Fameli, Roberta Nannucci: Legal Decisions and Integrated Systems. 291-296
- E. Jukes, J. Platts, S. Torrance: Legal Precedent Retrieval through Case-Based Reasoning and Semantic Interpretation. 297-304

Applications I

- M. K. Abdi, B. Kouninef, Maher K. Rahmouni: Towards a Case Tool for Information Systems Design. 305-314
- Steven A. Battle, Richard McClatchey: A Computerised Reservation System Using a Relational Database Augmented by Constraint Based Techniques. 315-321
- N. S. Binks, D. J. Smith: A Knowledge-Based System for Clinical Laboratory Result Interpretation and Validation. 322-328

Advanced Database and Information Systems Methods I

- Djamel Bouchaffra, Jean Guy Meunier: A Thematic Knowledge Extraction Modelling through a Markovian Random Field Approach. 329-338
- Bipin C. Desai: Internet and Searching Internet Resources. 339-349

Research Issues in Digital Libraries and Mobile Databases

- Aidong Zhang, Biao Cheng, Raj Acharya: Texture-based Image Retrieval in Image Database Systems. 349-356
- Robert P. Futrelle, Natalya Fridman Noy: Principles and Tools for Authoring Knowledge-Rich Documents. 357-362

Computer Supported Cooperative Work

- Igor Hawryszkiewycz: Knowledge Structures for CSCW. 363-369
- M. M. Ljubodraga, L. Janko: Control of Execution of Work Orders. 370-376

Physical Aspects

- Jérôme Besancenot, Michèle Cart, Jean Ferrié, Claire Morpain, Jean-François Pons, Philippe Pucheral: Preserving the Benefit of Strict 2-Phase Locking with Parallel Multidatabase Transactions. 377-386

Advanced Database and Information Systems Methods II

- Dushan Z. Badal, M. L. Davis: Investigations of Unstructured Text Indexing. 387-396
- N. Kamel, D. L. Hsiao-Feng: A Text Parser and Specification Generator for Modular ASN.1-Described Files. 397-406

Legal Systems II

A Thematic Knowledge Extraction in Text using a Markovian Random Field Approach

Djamel Bouchaffra and Jean Guy Meunier

Laboratoire d'Analyse Cognitive de l'Information (LACI)
University of Québec At Montréal (UQAM)
Centre ATOCI, Case postale 8888, succursale A,
Montréal (Québec), Canada H3C 3P8

bouchaff@pluton.atoci.uqam.ca
meunier@atoci.uqam.ca

Abstract.

We present a Markovian Random Field modeling for thematic knowledge extraction in text. An analogy is made between a flow of thematic investigations/textual fragments matching and statistical mechanics systems. The Markovian Field Knowledge Extraction machine (MAFKE) that we propose is based on a dynamical interaction between thematic queries and fragments composing a text. The representation of the textual knowledge system is submitted to state variations emerging from the flow of thematic queries. The MAFKE machine tries to satisfy the user thematic queries by changing the set of Units of Information (UNIFs) contained in a fragment. This change is computed with respect to the input thematic queries. Hence, MAFKE machine transists from one configuration state to another by changing the threshold assigned to the pertinency of UNIFs. For each state, a certain degradation of the system which depends on the thematic query index and this threshold is considered. The equivalence concept between an MRF and the Gibbs distribution (Max entropy) enables us to consider the energy and potential functions of this physical system. We use simulated annealing algorithm to isolate low energy states: this corresponds to the best (in some sense) knowledge extraction from the text that satisfies the user investigation. During the evolution towards these lower energy states, a fragment classifier emerges: the Markovian Random Field machine behaves as a classifier.

Index terms: Knowledge Extraction System (KES), Filtering, Markov Random Field, Gibbs Distribution, Knowledge System Degradation (KSD), Annealing, Classifier, Text Analysis, Informational Retrieval.

1. Introduction

Computer information management is confronted to data processing that comes more and more in the form of natural language text. And because this type of data is often very large, constantly growing and inquired in many ways, very little preprocessing (indexing, mark up, etc.) is possible. Hence, exploring, retrieving, analysing pertinent information becomes more and more difficult. Technical literature, be it from the point of view of an information retrieval [Salton et al., 1994 ; Burr, 1987] or contents analysis [Delany et al., 1993] shows that it has become urgent to develop tools to assist the creative exploration of large textual data banks. But one of the constraint on these tools is that they must respect the *dynamicity* (interaction with the user) and the *plasticity* (constant modification and proliferation) of a textual corpus .

From a methodological point of view, the intelligent processing of information in such a context encounters an epistemological problem. Indeed, one has difficulty with the AI postulate that calls upon knowledge representation for intelligent information processing. It is said that it is through this knowledgebe which can be either *structural* (syntactic, semantic, inferential, etc.) or *encyclopaedic* (objects, relations, events, situations, etc.) that an AI system could for istance, answer questions, retrieve pertinent sections, navigate in the text if not analyze and even "understand" it. Many researches have indeed shown that with a good knowledge base, there exist systems that can realize these types of tasks [Schank et al., 1994 ; Sowa, 1991 ; Recoczei et al., 1988 ; Jacobs et al., 1988 ; Moulin et al., 1990 ; Zari, 1990 ; Salton et al., 1994 ; Sabbah, 1989]. But it should be noted that these systems are often successful because they operate on familiar, relatively small, domain specific and well known texts.

In the situation of huge textual corpora where a processing system possesses only general structural kowledge, it becomes difficult to give before hand this system a pertinent and rich encyclopaedic knowledge. Indeed, how can one put in the knowledge base the new and specific knowledge that normally comes out of reading the text itself. One is placed here in the similar situation of the classical frame problem of AI.

In traditional AI expert systems, part of this question relates to the knowledge acquisition problem. In the design of these systems, the acquisition is done through some cognitive inquiry (protocol analysis) with the expert. But in a textual horizon, the expert knowledge is to be acquired through the processing of the text itself. The expert knowledge is in the text. So, how then is it possible to extract with a computer this knowledge from the text ? extraction that could then be used to built an eventual encyclopaedic knowledge base on specific themes.

Various mathematical models have been offered to deal with this problem. The more classical ones where inspired by various statistical pattern recognition or classicifation approaches such as clustering [Croft, 1980 ; Diday, 1987], component and factorial analysis etc., [Lebart et al., 1994 ; Cheeseman et al., 1988]. The problem with these models is their lack of sensiblity to the

dynamicity and plasticity of the inquiring situation. Computations has to be done anew for each modification of the data and for practically each type of request. Other, more recent models are inspired from neural networks such as the auto associative non supervised ones [Kohonen, 1982]. Although these models are sensitive to some aspect to dynamicity only but a few [Grossberg et al., 1987] present the plasticity required. Besides, hidden Markov models have been proposed, they revealed some auto-organization capabilities in a textual environement, however, the spatial intereaction inherent to these models is not very rich.

In this paper, we will explore a mathematical *model* that seems well fitted to the dynamicity and plasticity of the situation. More so, it seems also fitted to the incremental and interactive extraction of knowledge in a large textual corpus. We shall apply it to the specific problem of thematic analysis in text. Thematic analysis is a special knowledge extraction process. It starts by a first simple natural language request on the text. The request then is slowly modified by ajusting itself to the retrievied answers and develops into a full blown content analysis around some set of specific concepts that builds a theme. We propose to see the thematic inquiry as a dynamic and plastic interaction of a user with the corpus and to model it though the Markovian Random Field mathematical theory.

2. A dynamic and flexible knowledge acquisition system

The Markovian Field Knowledge Extraction machine (MAFKE) is an open architecture and a dynamical and flexible (plasticity) system.

(i) It is an open system because the objects that are the input of the *machine* are composed of entities (atoms) that can be of various semiotic types.

We call these latter entities UNITS of INFORMATION (UNIFs) [Meunier et al., 1993]. They can be linguistic units (words uniterm, complex terms lemmatized, whole sentence, etc.) or they can be some specific constituants of visual objects (such as pixels, measured by the grey level intensity, etc.) or physical signals (curves characterized by their frequencies, their amplitudes, etc.), a patient in a medical file (charaterized by the symptoms associated to classes of diseases) and iconic objects (such as a file icon, an application icon, etc.). Each UNIF is identified by some (manual or automatic) processes. In the following application the UNIFs will be defined as the set of linguistic word sequences of a text. All corpus fragments are built from such units of information. In other words, a fragment is a part of text and includes an arbitrary number of such units of information. For simplicity reasons, we shall write "unifs" instead of "UNIFs".

(ii) The system is said to be *dynamical* because of the constant interactivity between the user and the machine.

This is mathematically modelled by the mutual influence between a filtering process and an inquiring process that constantly modifies the states of the machine.

(iii) The system it is said to be *flexible* or plastic because it is not constrained by the length, the number and the types of input. This means that it can deal with corpora that are in constant modification.

2.1. The general schema of MAFKE

The essential characteristics of the MAFKE model is that it imposes a dynamical transformation of the state of the information. This transformation is obtained through a series of operations. Firstly, when a query is presented, it imposes a process that in turn, selects the set of unifs of a fragment that are the objects of the thematic inquiry. Secondly, it modifies through a series of proximity measure based on the density (with respect to the unifs) of each fragment of the corpus. Thirdly, it modifies through some similarity measure the clustering of the fragments and defines some subclusters on them (cliques). Fourthly, it enters in an interactive modification of the state of information through a series of queries the user will impose on the corpus. Each time, some parameters are modified and therefore, the configuration of the information in the data base of the system is in a constant mutation. This process is reiterated in a dynamical fashion and is modelized through the Markovian random field theory. This is done until a certain stability is attained.

2.2. Description of the different MAFKE components

We shall now present in a more formal way the MAFKE system.

2.2.1. The knowledge fund

In our approach, the knowledge fund is a full text written in natural language. This constitutes the *original* text. It is in a constant augmentation as the corpus is processed. This original text is subdivided into *fragments*. A Fragment is it self composed of many words (or unifs), and each unifs can be of any length and of various types. In our application, the unifs are either words, phrases, or even full sentences or a paragraph. For example, here is a set of two fragments containing all three sequences of unifs which are sentences:

Fragment 1

Unif 1 : Children in certain countries do not have the same rights as adults have.
Unif 2: In these countries, adults have indeed privileges than children don't have.
Unif 3: Often men are not always aware of the power they have over children.

Fragment 2

Unif 1: In many third world countries, poverty often hits the children more than the adults.
Unif 2: They are often left to themself, either for food or shelters.
Unif 3: The Unesco institution has a mission towards these children.

2.2.2. The thematic inquiry

A thematic inquiry is a contents research done on the corpus from a specific theme or set of concepts. This theme is expressed in natural language sentence query (affirmative or interrogative). Often this query is expressed as a single unif but it may also contains more than one. For instance, in our example one could have a research done of the theme of children rights. It could be expressed as:*" What are the right of children in the various countries of the world? "*
Here one could be looking in the text for the various statements that relate to this theme. The set of statement found will constitute the conceptual network of the theme under inquiry. In our model, we shall consider a query as a search for the presence or absence of hypothetical fragment in the knowedge fund.

2.2.3.. The MAFKE filtering operations

A filtering based on three operations is executed in the MAFKE machine: the reduction operation φ_1, the weighting φ_2 and the selection φ_3.

2.2.3.1. The reduction operation

It is a lexical discrimination applied in each fragment of the corpus. Using a dictionary, this operation eliminates a certain number of uniterms or complex words. Some sentences (unifs) which are not pertinent (in some sense) are also eliminated in this step. To apply this discrimination between unifs (smaller or larger), some techniques may exist in the literature, however, this remains an other interesting part of research which is not our objective at the present time.

2.2.3.2. The weighting operation

The first step consists to compute weights associated to the remaining atomic objects which are the uniterms and complex words. Several methods based on frequencies distribution exist [Salton, 1989]. Some of them construct a pertinency function based on the proximity between the statistical model (frequency, standard deviation, etc.) and the multinomial theoretical model. Hence, a weight proportional to this model proximity is assigned to each uniterm or complex term: it is *the first weighting operation.* The second step is to induce the previous weighting process to *a whole sentence* which is considered as an original unif. Hence, we have applied two operations: *a condensation* and *the second weighting operation.* Finally, one obtains a set of original unifs with their corresponding weights p_i in each fragment.

2.2.3.3. The selection operation

This last step represents a part of the dynamicity and plasticity of the MAFKE machine. Indeed, among the original unifs existing in each fragment, we select the unifs whose weights belong to the interval $[\zeta..\zeta+h]$ where ζ is a real variable representing a weight threshold called also a pertinency threshold and h is the length of the interval. Finally, one obtains a subset of the original unifs (sentences) with their corresponding weights. These unifs are those which are in "direct" interaction with the inquirer.

We shall present here in a graphical manner the preceding operations.

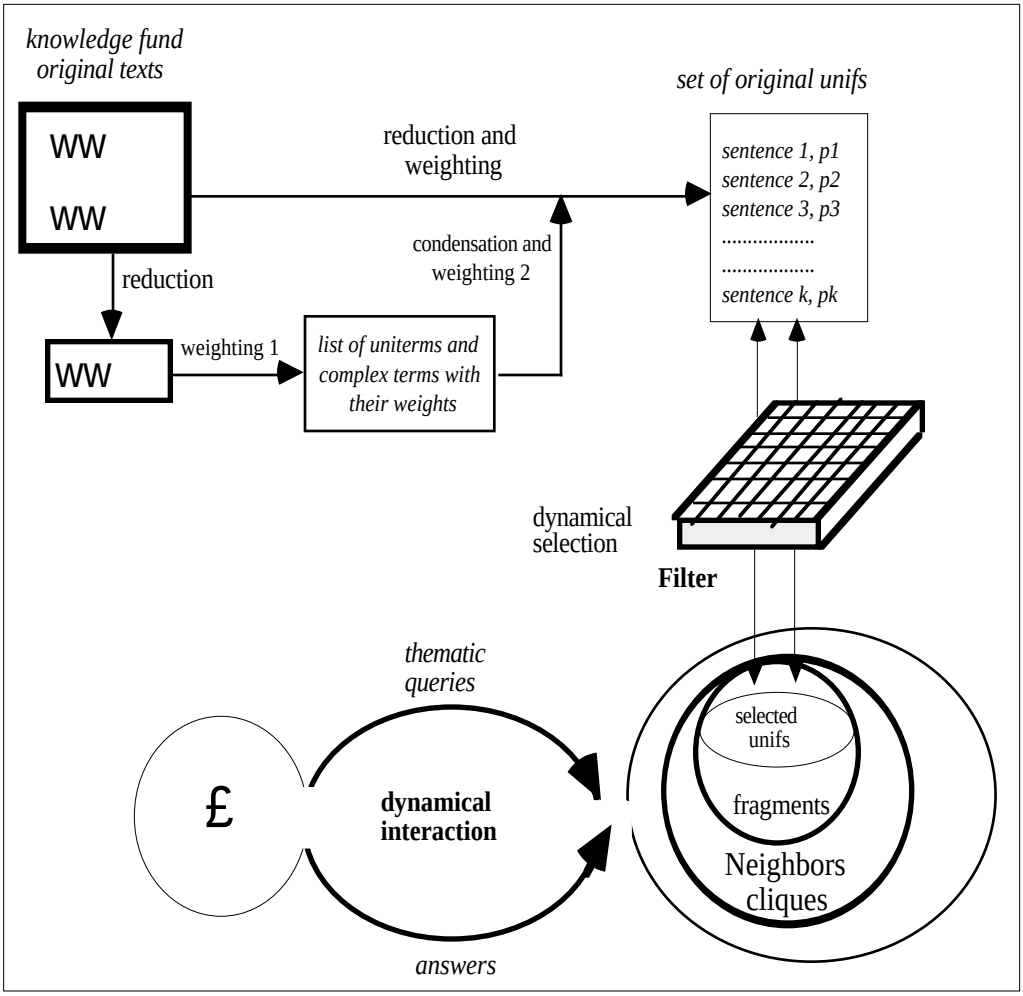

Fig. 1. A dynamic view of a Knowledge Extraction system.

In this global view, the MAFKE machine is a very interactive system. So much, so that it is not the investigator who at first modifies his queries from answers provided by the machine but on the contrary: it is the system that first modifies the state of information configuration in order to answer to the thematic queries of the investigator. In other words, it is the system that transforms the states of the information data in order to satify the interrogator. Through time, a learning processs develops[1]. The interrogator passes from a "naive" state to a "learned" state. He acquires more and more knowledge of what is in the fund.

[1] Other approachs to machine learning system have been conceptualized, see for example [Mitchell et al., 1986].

2.2.4. A morpho-syntactic distance and the compactness analysis

After having produced a set of fragment with a selected set of unifs, the MAFKE machine now starts operating on the unifs of the fragment. It will try to measure the proximity between the unifs in a fragment (containing the query considered as a hypothetic unif). More precisely, it tries to see how close or distant they are from each other. This is done by the comparison of certain linguistic features of the unifs. Although not all models of Knowledge Extraction believes in the effectiveness of linguistic analysis [Salton, 1989], we think that some morpho-syntactic analysis is pertinent in a dynamical system of Knowledge Extraction. The summarized approach that we present here is based on a paper of Pêcheux [23]. Pêcheux has defined a numerical strategy for calculating the distance between two word sequences based on a proximity of their various morphological forms of words. We shall apply this idea to evaluate the distance between a query and a unif that so far are considered both as two sentences[2]. This proximity measure between a query and a unif is based on a cost function associated to part of speech of each word composing a sentence and to a series of "edition" operations. This explains why we call this measure a morpho-syntactic distance. The elementary "edition" operations are: deletion, substitution and insertion of a character in a word. These operations are defined on a vocabulary V. For each elementary operation s, a positive number $\gamma(s)$ representing a cost associated to this opeartion is computed.

Example:

Query: ***"what are the right of children in the various countries of the world ?"***
Unif 1: *"children in certain countries do not have the same right as adults have."*
Unif 2: *"in these countries, adults have indeed privileges than children don't have."*

Using this distance, the result will be "the unif 1 is closer than the unif 2 to the query".

2.2.4.1. The edition distance between two words U and V

Definition 2.2.4.1.1. The "edition" distance between a couple of words $(U,V) \in V^*$ (free monoid generated by the vocabulary set V) is defined as:

$$\gamma^*(U,V) = \operatorname{Min}_{s \in \Delta} \gamma(s) ,$$

where Δ is the set of elementary operation series which transform the word U into the word V, (example: loved ---> love). The proximity measure γ^* is a distance in the topological sense.

[2] The choice of a sentence as a particular case is not a constraint inherent to the method.

2.2.4.2. A distance between two sentences P and T

We now can measure the distance between two sentences (or unifs). The similarity between sentences uses two different costs, the first one is based on weights given to parts of speech assigned to words and the second one is based on the words themselves composing the sentences.

Let P = p_1/r_1 p_2/r_2 ... p_k/r_k ... p_n/r_n and T = q_1/t_1 q_2/t_2 ... q_k/t_k ... q_m/t_m, two sentences constituted of two series of canonical words p (n words) and q (m words), (r_i), (i = 1,n) and t_j (j = 1,m) are their parts of speech provided by a lemmatizer.

Definition 2.2.4.2.1. The distance between two sentences P and T is a mapping from the cartesian product (S*S) (S: sentences space) to positive real numbers set (IR^+) and it is given by:

$$I(P,Q) = \frac{\gamma^*(P,Q)}{N} ,$$

where N is a normalizing factor dealing with the parts of speech, it is defined as:

$$N = \sum_{k=1}^{k=n} C_{d,i}(r_k) + \sum_{k=1}^{k=m} C_{d,i}(t_k) ,$$

and $C_{d,i}$ is the cost assigned to deletion and insertion elementary operations applied to the parts of speech. For more information about this morpho-syntactic distance, see [25]. In summary, this whole operation gives a proximity between two sentences. Thus, we have a distance between unifs contained in a fragment. An internal representation module can transform an inquiry into a standard form when it is necessary. This is very useful when one has to compute the proximity measure between a paraphrased query and a unif. The same remark can be stated in the case of an elliptic situation.

3. The neighborhood knowledge system

The MAFKE machine has measured the distance between unifs in each fragment. Now, it will measure the proximity between fragments in the corpus. This proximity defines a neighborhood system in the knowledge fund.

3.1. The neighborhood system between textual fragments

In our approach, the configurations of the information in the knowledge fund is seen as a set of fragments forming a neighborhood system based on a certain similarity between them. This similarity is measured through some type of proximity. The neighborhood system is inherent to the use of a Markovian Random Field process.

In more formal terms, a neighborhood system can be defined in the following manner.

Definition 3.1.1. Let $\mathcal{J} = \{f_1, f_2, ..., f_k\}$ be a set of fragments (vertices in a graph), $V = \{V_f, f \in \mathcal{J}\}$ is called a neighborhood system for $\mathcal{J}$ if it is a subset of $\mathcal{J}$ such that:

$$\begin{cases} f_i \notin V_{f_i} \\ f_i \in V_{f_k} \Leftrightarrow f_k \in V_{f_i} \end{cases}.$$

The doublet $(\mathcal{J},V)$ is a hypergraph of order k = card($\mathcal{J}$) where a hyperedge is composed of all fragments which are neighbors according to some sense.

Before constructing the neighbors set, we use a proximity measure between two fragments called "the Mean distance" which is very useful when the distance is not Euclidean, it is introduced in the following definition:

Definition 3.1.2. If I is the morpho-syntactic distance between two unifs, then a "Mean distance" between two fragments can be given as:

$$d_M(\text{fragment u, fragment v}) = \frac{1}{P_{\text{fragment u}} \cdot P_{\text{fragment v}}} \sum_i \sum_j I(\text{unif i}, \text{unif j}),$$

where "M" stands for the "Mean" and the couple (unif i,unif j) belongs to the cartesian product ({fragment u} * {fragment v}). The values $P_{\text{fragment u}}$ and $P_{\text{fragment v}}$ are the global weights associated to each fragment. They can be written as:

$$\begin{cases} P_{\text{fragment u}} = \sum_i P_{\text{unif i}} \cdot \chi_{\text{fragment u}}(\text{unif i}) \\ P_{\text{fragment v}} = \sum_j P_{\text{unif j}} \cdot \chi_{\text{fragment v}}(\text{unif j}) \end{cases},$$

where χ is the characteristic function defined as:

$$\chi_{\text{fragment u}}(\text{unif v}) = \begin{cases} 1 \text{ if unif v} \in \text{fragment u} \\ 0 \text{ otherwise} \end{cases} .$$

One can find other types of nonEuclidean distances between two groups of objects in the Cluster Analysis community [Anderberg, 1973 ; Lebart et al., 1994]. The neighborhood system is based on a proximity measure between objects of the same type. That is to say, the proximity measure d_M is applied with respect to the sets of unifs contained in two neighbor fragments. The neighborhood system can be written as:

$$V(\alpha)_{\text{fragment i}} = \{\text{fragment j} ; d_M(\text{fragment i}, \text{fragment j}) \leq \alpha, \\ \alpha > 0 \text{ is a neighborhood threshold}\}.$$

Very often, when we want to identify some differences between groups, one expresses a degree of discrimination between these groups by using the inertia concept (variance within and between fragments). As outlined previously, this is possible when we are in presence of an Euclidean distance as it is the case for example when one projects the fragments space into the vectorial weights space [Salton et al., 1994], i.e., a fragment is replaced by the weights assigned to its unifs. However, we believe that the unifs space is not necessarily Euclidean, this is the reason why we adopt "the distance of the mean" between two fragments as a base of the neighboring system construction.

Definition 3.1.3. For a fixed value of α, $V(\alpha)_{\text{fragment i}}$ is called a neighborhood configuration associated to fragment i.

In summary, this operation allows the MAFKE machine to form clusters of fragments that have some type of similarity among them.

Definition 3.1.4. A subset C $\subseteq$] is called a clique of fragments if every pair of distinct fragments in C are neighbors. With respect to graph theory, the clique order here is equal 2.

This latter concept enables the MAFKE machine to extract a special subcluster, called cliques. These cliques aims to identify more closely related fragments in a particular neighborhood configuration.

3.2. The thematic inquiry / fragment matching

3.2.1. The thematic query projection

As outlined above, a thematic query can be interpreted as an investigation to compare a hypothetical fragment with the set of fragments present in the knowledge system. If so, then one can apply the preceding proximity measure between the thematic query and a fragment. However, the structural nature of a thematic query is more akin to the one of a unif than to the one of a fragment. In other words a thematif query can be considered as a particular unif. We project the thematic query τ in the set F of unifs associated to each fragment (Fig 2a). Then, we compute the compactness and standard deviation parameters associated to the set $F \cup \tau$. Hence, depending on the values of a function based on compactness and standard deviation parameters, one can induce a proximity measure between a thematic inquiry and a fragment. Thus, this proximity measure gathers the fragments in different groups (neighboring configurations). However, the MAFKE objective is to match the thematic inquiry with each group of fragments (see Fig2b) and extract different cliques from this group.

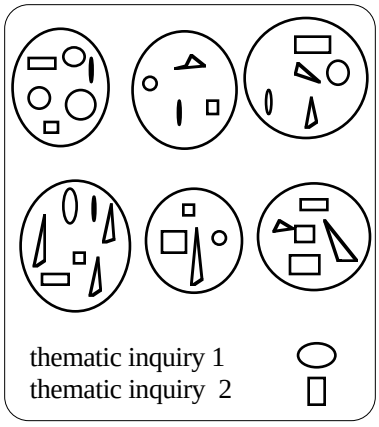

Fig. 2a. Fragments containing unifs.

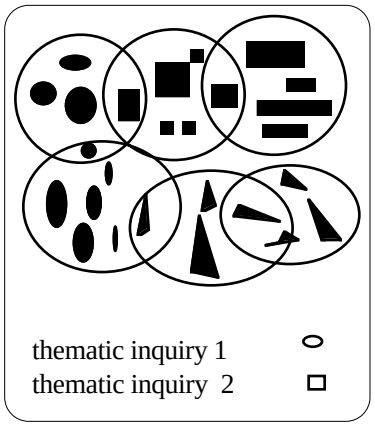

Fig. 2b. Neighborhood configurations between fragments.

3.2.2. The fragment compactness variability and the user decision

In order to extract the knowledge that may satisfy the interrogator investigations, the MAFKE machine in response to a user thematic query proposes at the begining of the investigation the fragments with low compactness (density). However, as the number of queries / answers between the user and the machine increases and reachs an ideasyncratic beta point (see figure 3a), the chance of selecting by the machine fragments with high compactness will increase. In other words, the machine gets more and more informations about the investigations undertaken by the user. The information precision depends on the quality of the Markovian Random field model supporting the machine. Thus, at the end of the queries / answers flow, the MAFKE machine may focus on the interrogator desired knowledge. Formally, this can be described by the following definitions:

Definition 3.2.2.1. The degree of extracting fragments responding to a thematic query execution is represented by a positive real function value called: the query / fragment "matching".

Definition 3.2.2.2. The matching function is a mapping $M_{f,\tau}$ from $F^*\{\tau\}$to positive real numbers with the following analytical structure:

$$F^*\{\tau\} \xrightarrow[M_{f,\tau}]{} IR^+$$

$$(unif_1, unif_2, ..., unif_n\ ;\ \tau) \xrightarrow[M_{f,\tau}]{} M_{f,\tau}(NBIT, \bar{I}_f) = \frac{(\sigma_f(I)\ .\bar{I}_f - NBIT^{\beta})^2}{2a^2} + \frac{(NBIT + \bar{I}_f^{\beta})^2}{2b^2}\ ,$$

where $\bar{I}_f$ and $\sigma_f(I)$ are respectively defined as:

$$\bar{I}_f = \frac{\sum_{k=1}^{k=(n+1)} \sum_{j=(k+1)}^{j=(n+1)} w_{k,j} I_{k,j}}{\binom{2}{n+1}} \quad ; \quad \sigma_f(I) = \left(\frac{\sum_{k=1}^{k=(n+1)} \sum_{j=(k+1)}^{j=(n+1)} w_{k,j} (I_{k,j} - \bar{I}_f)^2}{\binom{2}{n+1}} \right)^{\frac{1}{2}} .$$

The set $F \subseteq G$ corresponds to the n unifs contained in a fragment, the term (n+1) comes from the fact that we project the query into the fragment of n unifs, τ is the thematic query. The parameter "NBIT" is the number of iterations (an iteration is a couple (query index, threshold value ζ_i)), $w_{k,j}$ [3] are weights associated to each couple of unifs and they are supposed to be equal, I is the morpho-syntactic distance between unifs. The matching function surface (figure 3a) can be approximated by *an elliptic paraboloid.*

[3] Different weights can be assigned to distinguish between the morpho-syntactic connection strengths of couples of unifs.

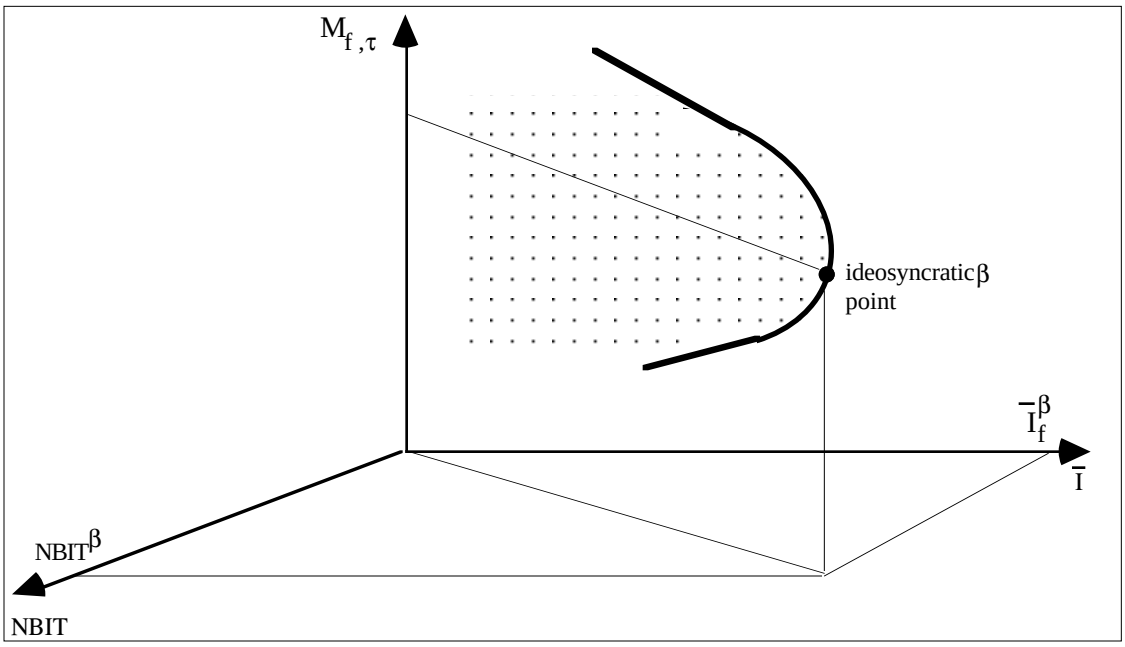

Fig. 3a. The query-fragments matching function in 3D.

The β point corresponds to the user decision time to focus on compact fragments (small means and standard deviations) extracted from the knowledge fund. The figure 3b shows the evolution of a series of fragments that can be also decided by a user.

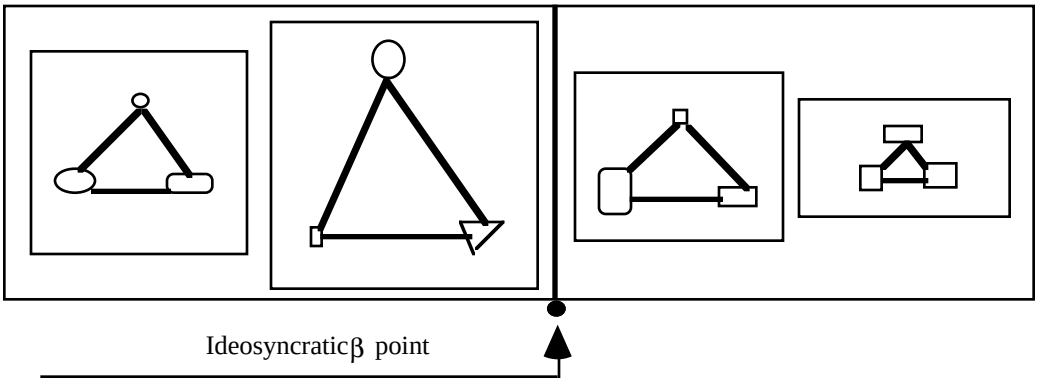

Fig. 3b. The vertical bar represents a "focalization point" on fragments decided by the inquirer.

4. The Markov Random Field concept in a KES

4.1. The thematic query / fragment MRF

It seems much more consistent to study the query / fragment matching not in all the knowledge fund but only among a packet of fragments which are neighbors (according to the sense defined in section 3). For example, instead of covering all the fragments representing the knowledge fund which is useless and much more expensive in cpu time, one has to cover only some particular set of neighboring fragments.

Definition 4.1.1. The query / fragment matching $M_{f,\tau}$ is a stochastic function and it is considered as a Markov Random field $X_{f,\tau}$ with respect to V.

Let $X = \{X_f, f \in \mathcal{J}\}$ denote any family of random variables indexed by $\mathcal{J}$ (set of fragments contained in the corpus). Let Ω be the set of all possible configurations:
$\Omega = \{\omega = (x_{f1},x_{f2},...,x_{fk})\}$, $f_i \in \mathcal{J}$ and $i \in [1..k]$. If one abbreviates $\{X_{f1}= x_{f1},...,X_{fk} = x_{fk}\}$ as $\{X = \omega\}$ then it follows:

Definition 4.1.2. The variable X is *a Markov Random Field* with respect to a neighborhood system $\{\mathcal{J},V\}$ if:
Prob $\{X= \omega\} > 0$ for all $\omega \in \Omega$
Prob $\{X_f = x_f / X_r = x_r, r \neq f\}$ = Prob $\{X_f = x_f / X_r = x_r, r \in V(\alpha)_f\}$

The definition of Markovian Fields by means of conditional probabilities has been first proposed by Dobrushin [10]. In our approach, the MRF expresses the local interaction between fragments (f_j), $j \in [1..k]$ with respect to the user thematic query τ. In other words, for a fixed query τ, and a fixed fragment f_p, when one knows the matching degree of this query on all fragments of the fund except the fragment f_p, then the information about the matching degree of the fragment with the thematic query depends only on fragments which are neighbors to fragment f_p. Formally, this can be stated as:

$$\text{Prob}\{X_{f_p}(\tau) = x_{f_p} \;/\; X_{f_q}(\tau) = (x_{f_q})_{q\neq p}\} =$$
$$\text{Prob}\{X_{f_p}(\tau) = x_{f_p} \;/\; X_{f_r}(\tau) = (x_{f_r}) \,;\, f_r \in V_\alpha(f_p)\}\,.$$

We consider that there is a spatial interaction between fragments with repect to two different types of slow variables which are: the number of queries / answers between the user and the system and the pertinency threshold ζ (slower variable) assigned to each unif in the system.

Definition 4.1.3. A configuration $\omega = (x_{f_1}(\tau,\zeta), x_{f_2}(\tau,\zeta), \ldots, x_{f_k}(\tau,\zeta))$ of the system is a set of k realizations (k is the number of fragments in the corpus) of the MRF X.

The contribution of Hammersley-Clifford enables us to consider and have a vision of a quantitative state of all fragments simultaneously (joint probability) but not only separately [Besag, 1974]. Besides, this state can be seen as a state of particles in statistical mechanics environment. The following equivalence theorem allowing this passage can be formulated as:

Theorem 4.1.4. *[Hammersley-Clifford] The random variable X is a Markov Random Field with respect to a neighborhood system {I,V} if and only if Prob(ω) = Prob{X = ω} is a Gibbs distribution with respect to the same neighborhood {I,V}.*
Proof. See [Kindermann et al., 1980].

4.2. Why an MRF approach?

The hypotheses of using the concept of a Markovian Field in describing the group of similar fragments with respect to a flow of thematic queries in a knowledge fund originates from the inadequacy of most classical information retrieval methods [Ellis, 1990] to take into account the dynamicity involved in the local interaction that must go on between an information base and requests from users. The problem is even more evident in a thematic analysis that calls upon a constant interaction between the queries and the textual corpus. As outlined at the begining, in the classical models the clustering imposed or discovered on the textual fund is stable and does not change with respect to the inquiry. The thematic queries may be change but the configuration of the information in the knowledge fund does not. This is a major problem for thematic inquiries, a theme is not explored by only receiving a fixed set of answers to a single query but to a set of queries that themselves are modified throughout the process which in turn modifies the vision of the information in the knowedge fund. This implies a strong dynamicity between the knowledge fund and the queries. The MAFKE hypotheses models this process. It allows to describe the influence a thematic query may have on the configuration of the information in the knowledge fund. Technically, this is done by identifying how the neighborhood relations among fragments are influenced with respect to a flow of thematic queries. A series of queries may regroup or separate the fragments differently. This latter move provides a change in the neighborhood configurations (hyperedge) oganization. Hence, the configuration of groups of fragments are in constant change due to the introduction of the thematic queries. Besides, we shall see in the complete version of this article that the theory of Markov

Random Field uses the notion of energy assigned to a configuration of the system. Instead of using probabilities that are unknown, one has to use potentiel functions [Dobrushin et al., 1993] derived from energy which is easier. Here, we consider only the pair potentials which are *translation invariant isotropic*. We are experimenting some of them (Morse, etc.) having the following property: *two fragments can meet together by a user query*.

5. The knowledge extraction degradation

Our aim here is to provide a mathematical model capable to simulate the dynamics of the query configuration of the informations in the fund. In fact, if we translate this dynamicity into energy terms, one could see the knowledge fund through the interaction of the flow of queries and the modification of neighborhood configurations arrive at a stable state (as the majority of the gradient dynamics). At this attractor state, modification of configuration become less and less possible. We say then that the energy in the knowledge fund is at a low level of degradation (low energy state). The knowledge interaction system depends at least on the accuracy of the unknown threshold ξ associated to the functions φ_3, the set of asynchronous queries and the analytical structure of the morpho-syntactic distance I. For example, the distance I between a query and an unif can be corrupted by morphological and syntactic ambiguities assigned to them (different trees parsing, etc.), [Bouchaffra et al., 1993, 1994]. The distance I, the query τ and ζ are involved in the definition of the Markov Random Field X. However, their roles with respect to the system can be divided into two classes: The class of the internal variables (fast variables) and the class of exterior variables (slow variables) to the system. Besides, an other parameter which is inherent to the query structural nature is the degree of relevance of queries emanated from the user with respect to the theme investigated: this can be considered as: the competence level of the user with respect to the theme explored. Before the user begins his inquiry, the interaction of the knowledge extraction system is considered as degraded because there is no specific available knowledge that can be extracted at this time. Hence, one can decompose this degradation into the following manner:

$$\textit{Total degradation} = \textit{system degradation} \otimes \textit{user degradation},$$

where '$\otimes$' is an invertible operator as '+' or 'x' etc.

Remark 5.1. The matching degree of fragments by a flow of thematic queries is not necessarily uniform in the whole knowledge fund, it very often depends on the theme encountered. It can be weighted equally with the average of the nearest neighbor fragments.

This remark obliges us to consider a weight function W which remains constant in a neighborhood configuration of fragments. Hence, for a fixed ζ_i and a fixed inquiry τ_j, one can express analytically the nonlinear degradation modeling of the representation of the knowledge fund B as:

$$Deg^{\zeta i}(\tau_j) = \psi\,(W(X^{\zeta i}(\tau_j))) \otimes N^{\zeta i}(\tau_j),$$

where $\psi(.)$ is a nonlinear invertible function as (x ---> ln(x) or x---> sqrt(x)).
Analytically, this can be written as:

$$d^{\zeta_i}_{\tau_j,\,B} = \psi\left(\sum_v w_{\tau_j,\ \text{fragment v}} \cdot x^{\zeta_i}_{\tau_j,\ \text{fragment v}}\right) \otimes n^{\zeta_i}_{\tau_j,\,B} \quad \forall j\ ,$$

where w is defined as:

$$w_{\tau_j,\,\text{fragment u}} = w_{\tau_j,\,\text{fragment v}} \ \text{ if } \ \text{fragment v} \in V_\alpha(\text{fragment u})\ .$$

We also decompose the user degradation with respect to the neighborhood configurations as following:

$$n^{\zeta_i}_{\tau_j,B} = \sum_k \lambda_{NbC_k} . n^{\zeta_i}_{\tau_j,NbC_k}\ ,$$

where NbC stands for a neighborhood configuration, the random noise $n^{\zeta_i}_{\tau_j,NbC_k}$ represents the random flow of queries emanated from the user, it is supposed to be Gaussian with mean μ and variance σ^2. The parameters λ_{NbC_k} are weight coefficients associated to each neighborhood configuration.

Remark 5.2. For each couple (ζ_i,τ_j) we have a specific degradation, the MAFKE machine transists from a degradation state to another.

6. The answers from the machine

When an investigator submits a query τ to the machine assigned to a certain thershold ζ, different values of the Markovian field $x_{f,\tau}$ (NBIT) of fragments are computed. We then have a configuration ω = X(NBIT) = ($x_{f1,\tau(}$NBIT), $x_{f2,\tau}$(NBIT) ,..., $x_{fk,\tau}$(NBIT)) of all fragments in the fund with respect to the query τ and the threshold ζ. We selects fragments as answers to the investigator by considering the local characteristics of the conditional Gibssian distribution. In other words, we

choose un certain number of fragments from the conditional distribution of X given the observed values $x_{f,\tau}(NBIT)$ of the neighboring fragments $x_{s,\tau}(NBIT-1)$ where $s \in V(\alpha)_f$. Given an initial configuration of the fund X(0), we thus obtain a series (X(NBIT)) (NBIT $\in$ A (subset) $\subset$ IN (natural numbers)) of configurations which evolves due to the interaction change between queries and the knowledge fund that converges to a certain limit state which does not depend on X(0). The description of the algorithm supporting this answers process is similar to the well known "Gibbs Sampler".

7. The optimal (Bayesian) configuration

Our aim is to determine an optimal configuration $\omega^* = (\hat{x}_{f_1}, \hat{x}_{f_2}, ..., \hat{x}_{f_k})$ and an optimal threshold ζ^* with respect to a flow of thematic queries. Formally, the problem can be written as[4] :

$$\underset{\zeta,\omega}{Max}\left\{Prob\,(X^{\zeta} = \omega / d^{\zeta})\right\} \Leftrightarrow \underset{\zeta,\omega}{Min}\left\{E_p^{\zeta}(\omega)\right\},$$

where " E_p^{ζ} " stands for posterior energy function which depends on the degradation d^{ζ}.

This probability is a Gibbs distribution and the problem is transformed into the research of a configuration ω^* and a threshold ζ^* at a minimal energy state. By decreasing the temperature parameter T depending on two slow variables which are the threshold ζ and the index query τ, we reach configuration states (fast variables) of lower energy. The temperature parameter T is a nonlinear function converging to zero when the number of iterations increase. In order to avoid local minima, we tolerate with a certain probability higher energy states: it is the simulated annealing algorithm [Kirkpatrick et al., 1982] that we apply.

8. Conclusion

As outlined above, thematic analysis is a special knowledge extraction process. It is not a retrieval of document pertinent to a query but a heuristic exploration of a theme in a corpus of which one ignore the contents. The model we have proposed in this paper allows a user to explore in an intelligent way a textual corpus. An inquirer, that has a theme to be explored, will start by a single natural language question. This question will be matched to the transformed corpus but in an unusual manner. Because of the Markovian Field property, it will affect the configuration of the information

[4] The equivalence between these two optimisation problems is due to the posterior Hammersely Clifford theorem and it is shown in the complete version of this paper.

in the knowledge fund and find a first type of answers (fragments) accordingly. The answer will in turn influence the inquirer that will reformulate his query. This process will go on until no more transformation of the query or the configuration of the database is attained. That is, no more significant modification of the clusters and of the query is possible. As a result, the inquirer will have in hand a whole set of answers from which he will build his own interpretation of the theme he had originally in mind. We believed that this model can help structuring the technical problem of knoweldge extraction from text. Indeed, because a text cannot be read before it is processed, it is difficult to build in the extraction machine the encyclopaedic knowledge specific to the text under scrutiny. This means that knowedge extraction is dependent on the dynamic interaction that will take place between the queries and the knowledge fund itself. This interaction is ideosyncratic to a user .

References

[1] M. R. Anderberg, "Cluster analysis for applications", *Academic Press, 1973.*

[2] J. Besag, "Spatial interaction and the statistical analysis of lattice systems, (with discussions)", *Journal of the Royal Statistical Society, JRSS, Series B, Vol. 36, n° 2, pp. 192-236, 1974.*

[3] D. Bouchaffra, G. Lallich-Boidin and J. Rouault, "stratified sampling with Bootstrap restimation: application to parts of speech tagging", *Secondes Journées Internationales d'Analyse Statistiques de Données Textuelles, Montpellier, France, 21-22 Octobre, 1993.*

[4] D. Bouchaffra and J. Rouault, "Different ways of capturing the observations in a Hidden Markov Model: application to Parts-of-Speech Tagging", *eds: P. Cheeseman and R.W. Oldford: Selecting Models from Data: Artificial Intelligence and & Statistics IV, Lecture Notes in Statistics Ser: Vol 89, Springer Verlag, 1994.*

[5] D. J. Burr, "Experiments with a connectionist text reader", *IEEE First International Conference on Neural Networks, San Diego, 1987.*

[6] P. Cheeseman, M. Self, J. Kelly, J. Stutz, W. Taylor and D. Freeman, "Bayesian classification", *Proceedings of AAAI, Minneapolis, 607-611, 1988.*

[7] W. B. Croft, "A model of cluster searching based on classification", *Information Systems, 189-195, 1980.*

[8] P. Delany and M. Landow, "The digital word: text based", *eds. MIT press,Computers and the Humanities, Cambridge, MA, 1993.*

[9] E. Diday, "Orders and overlapping clusters by pyramids", *Rapport de recherche n° 730, INRIA, 1987.*

[10] R.L. Dobrushin, "The description of a random field by means of conditional probabilities and conditions of its regularity", *Theory Prob. Appl., Vol.13, 1968.*

[11] R.L. Dobrushin and S. Kusuoka, "Statistical mechanics and fractals", *Lecture notes in mathematics, Springer Verlag, 1993.*

[12] D. Ellis, "New horizon in Information Retrieval", *London: Library Association, 1990.*

[13] S. Grossberg and S. Carpenter, "Self organization of stable category recognition codes for analog input patterns", *Applied optics, 26, 4919-4930, 1987.*

[14] P. Jacobs and U. Zernik, "Acquiring lexical knowledge from text: a case study", *Proceedings of AAAI 88, University of St Paul, Minesota, 1988.*

[15] T. Kohonen, "Clustering taxonomy and topological maps of patterns", *IEEE, Sixth international conference on pattern recognition, 114-122, 1982.*

[16] S. Kirkpatrick, C.D. Gellatt, Jr. and M.P. Vecchi, "Optimization by simulated annealing", *IBM Thomas J. Watson Research Center, Yorktown Height, NY, 1982.*

[17] R. Kindermann and L. Snell, "Markov Random Fields and their applications", *Contemporary mathematics, AMS, Vol. 1., 1980.*

[18] L. Lebart and A. Salem, "Statistique textuelle", *Dunod, Paris, 1994.*

[19] J. G. Meunier, S. Bertrand-Gastaldy and H.Lebel, "A call for Enhanced on text content for information retrieval", in International classification, pp. 5-15., 1987.

[20] J. G. Meunier, S. Bertrand-Gastaldy et L.C. Paquin, "La gestion et l'analyse de textes par ordinateur: leur spécificité dans le traitement de l'information", *en collaboration avec S. Bertrand-Gastaldy et L. C. Paquin, in Revue de Liaison de la recherche en informatique cognitive des organisations (ICO Québec), Vol. 6 1 et 2,pp. 19-41, printemps 1993.*

[21] T.M. Mitchell, J.G. Carbonell and R.S. Michalski, "Machine learning: A guide to current research", *Kluwer Academic, New York, 1986.*

[22] B. Moulin et D. Rousseau, "Un outil pour l'acquisition des connaissances à partir de textes prescriptifs", *ICO, Québec, 3, (2), 108-120, 1990.*

[23] M. Pêcheux, "Analyse automatique du discours", *Dunod, 1969.*

[24] S. Recoczei and P. Epo, "Creating the domain of discourse: Ontology and inventory", *In J & B.J. Boose (Ed.), knowledge acquisition tools for experts and novices, Academic press, 1988.*

[25] J. Rouault, "Analyse automatique du discours: Note sur le système 3AD75", *Communication interne, Université Stendhal, Grenoble, 1994.*

[26] G. Sabbah, "L'IA et le langage", *Paris, Hermes, 1989.*

[27] G. Salton, J. Allan and C. Buckley, "Automatic structuring and retrieval of large text file", *Communication of the ACM, 37, (2), 97-107, 1994.*

[28] G. Salton, "Automatic text processing: the transformation, Analysis and Retrieval of Information by Computer", *Reading, MA, Addison Wesley, 1989.*

[29] R. Schank and M.K. Colby, "Computer models of thought and language", *W.H. Freeman, 1994.*

[30] J.F. Sowa, "Principles of semantic networks", *San Mateo: Morgan, Kaufman, 1991.*

[31] G.P. Zari, "Représentation des connaissances sur des documents en language naturel", *In Office de la langue française (ed.), les industries de la langue, perspective 1990, Gouvernement du Québec.*