IEEE Xplore DIGITAL LIBRARY

SEARCH

Advanced Search | Preferences | Search Tips

IEEE

BROWSE | MY SETTINGS | CART | SIGN IN | About IEEE Xplore | Terms of Use | Feedback | Help

Browse > Conferences> Document Analysis and Recogni

Prev | Back to Results | Next »

A Markovian random field approach to information retrieval

Download Citation | Email | Print | Request Permissions

Bouchaffra, D.; Meunier, J.G.;
Lab. d'Analyse Cognitive de l'Inf., UQAM

Access The Full Text

SIGN IN: Full text access may be available with your subscription

User Name | •••••••• | Sign In

Forgot Username/Password? Athens/Shibboleth Sign In

Already Purchased? View Now. | Purchase Now

Not a subscriber?

Get full-text access with a subscription to the IEEE *Xplore*.

Which subscription is right for you?

LEARN MORE

This paper appears in: Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on
Issue Date : 14-16 Aug 1995
Volume : 2
On page(s): 997 - 1002 vol.2
Meeting Date : 14 Aug 1995-16 Aug 1995
Print ISBN: 0-8186-7128-9
Cited by : 1
INSPEC Accession Number: 5630595
Digital Object Identifier : 10.1109/ICDAR.1995.602070
Date of Current Version : 06 August 2002

ABSTRACT

A Markovian random field approach is proposed for automatic information retrieval in full text documents. We draw up an analogy between a flow of queries/document images connections and statistical mechanics systems. The Markovian flow process machine (MFP) models the interaction between queries and document images as a dynamical system. The MFP machine searches to fit the user's queries by changing the set of descriptors contained in the document images. There is hence a constant transformation of the informational states of the fund. For each state, a certain degradation of the system is considered. We use simulated annealing algorithm to isolate low energy states: this corresponds to the best "matching" in some sense between queries and images

A Markovian Random Field Approach to Information Retrieval

Djamel Bouchaffra and Jean Guy Meunier

Laboratoire d'Analyse Cognitive de l'Information, UQAM Canada
E-mails: bouchaff@pluton.atoci.uqam.ca ; meunier@atoci.uqam.ca

Abstract

A Markovian Random Field approach is proposed for Automatic Information Retrieval in full text documents. We draw up an analogy between a flow of queries/document images connections and statistical mechanics systems. The Markovian Flow Process Machine (MFP) models the interaction between queries and document images as a dynamical system. The MFP machine searches to fit the user's queries by changing the set of descriptors contained in the document images. There is hence a constant transformation of the informational states of the fund. For each state, a certain degradation of the system is considered. We use simulated annealing algorithm to isolate low energy states: this corresponds to the best "matching" in some sense between queries and images.

Index terms: Automatic Information Retrieval, filtering, morpho-syntactic distance, Markov Random Field, Gibbs distribution, annealing.

1. Introduction

Classical models of information retrieval (IR) usually present two implicit postulates. Firstly, the configuration of the representations (set of image descriptors) of a documentary fund is constant or stable [13,14]. Secondly, there exists one representation of the documentary fund that is useful for a variety of users. However, with the advent of large scale full text storage space and electronic highways, the documentary fund becomes more and more unstable and is processed by a mutliplicity of different users. It hence becomes very difficult to give to the fund a definite and a satisfactory representation. In this paper, we explore some alternative way to tackle the problem of a dynamic and a multiuser full text information retrieval system. Indeed, we approach it through a stochastic and dynamic model called Markovian Field Process (MFP). This model allows us to see the information retrieval process as a dynamical system. It is a system based on highly interactive relations between the user's queries and the fund's configurations. One associates a state sequence to this system and our aim is to reach steadily lower energy states. In this paper, we present the general and basic concepts of this Markovian Field Process and its application to the IR problem.

2. The drawbacks of classical IR methods

The classical frameworks of IR relies on two important processes. The first one is a filtering that extracts a set of image descriptors from an original document. The second one is a matching or comparison procedure between a query and images of the documents [4]. These two processes present particular difficulties.

(i) The first one relates to the fitlering process. In many models this filtering is done through a specific threshold value. This threshold value is often influenced by a theoretical model. In practice, it is based on some type of probability measure and it cannot be modified.

(ii) The second difficulty concerns the degree of certainty that can be credited to the similarity measure between a query and an image [16]. These similarity values are computed from weight vectors representing only a particular description of the fund. They are not sensitive to the descriptors that a user may include or value in his query.

(iii) The third difficulty is that in the classical retrieval model, it is difficult to take into account the updating of a documentary fund. The input of new documents in the fund affects all the documentary system objects and therefore the computation involved in these methods can become expensive in cpu time.

Some solutions using emergent approaches as neural networks have been suggested to these problems [1,18]. We believe that a Markovian random field machine or Markovian Flow Process (MFP) can answer some of the difficulties underlined above.

3. The framework of a dynamic I.R. system

We present a global view of the archtecture of the IR system.

3.1. The Markovian Flow Process (MFP) approach

The Markovian Flow process machine is an open architecture and a dynamic system.
(i) It is an *open* system in the sense that as input, it admits entities of various semiotic type that we call here UNITS of INFORMATION or UNIFs [10,11]. In the field of textual document, these UNIFs are usually linguistic units (such as uniterms, complex terms lemmatized, phrases, etc.). Each UNIF is identified by some (manual or automatic) processes, it can be of various length. All queries are built from such *UNIFs*.

(ii) The system is also seen as *dynamic* in the sense that it sees the interactions between filtering and quering as *energy states* of a dynamic machine. The MFP model approaches information retrieval as a dynamic transformation of the state of information in the system. This transformation is realized each time: a) a filtering process selects the descriptors from the predescritors original sets, and b) a query is done on the documentary fund (see figure1). Hence, the information state is in a constant mutation. In this view, it is not the interrogator 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. Through time, a learning processs develops. The interrogator passes from a "naive" state to a "learned "state. He acquires more and more knowledge of what is contained in the fund. Let us now present in the figure 1 the general schema of the MFP processing chain.

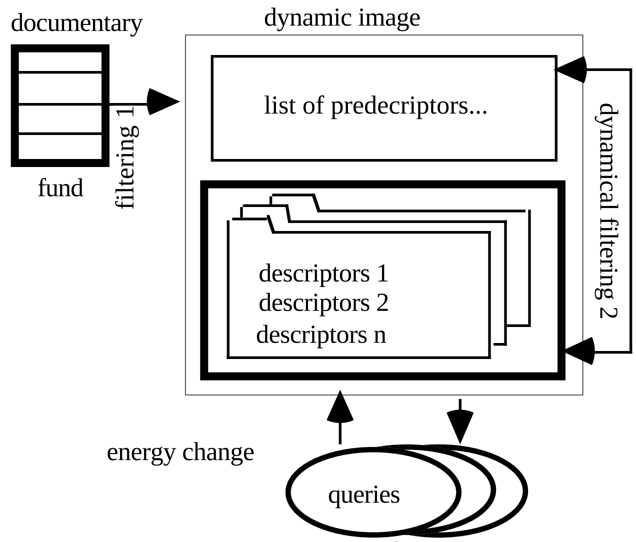

Fig. 1. A dynamic view of an IR system.

3.2. Description of the different MFP components

In our approach, the first component of the chain is the documentary fund. This is a full text data base written in natural language. It is subdivided into several fragments that are equivalent to individual document images. We define a fragment as a connected section of the documentary text. Each fragment of the documentary fund is constituted of a sequence of units of information. For example, here is one example of a fragment containing a sequence of units of information which is a sentence. this sentence is called a descriptor in our case.

Fragment. *The industry in the USSR is under an important pressure from the the Eastern Europe's industry*[1].

The second important component in the chain is the query. A query q_j can be seen as a hypothetical image. It is a question on the presence or absence of a hypothetical image in the documentary fund. So, it will also be composed of a sequence of UNIFs: it is a description of a hypothetical document. A query could be for instance:
"How does the industry in the USSR behave compare with the industry in Poland".

On these two components, two specific operations will be applied: a filtering process (filtering 1 + filtering 2) and a comparison procedure. The first step in the chain consists in building descriptors (constituting a dynamic image) out of the original textual fragments. These dynamic images are in a sense equivalent to images of classical IR but they are here not fixed and definite: they can change. These dynamic images are realized though two filtering processes. The filtering 1 process (that can be manual or automatic) consists in choosing a subset of unif sequences as constituting the "predescriptors" of the dynamic images of these fragments. In the case of a probabilistic filtering, the pertinency of a possible predescriptor may consider the scattering of this predescriptor in a fragment with respect to its dispersion in the whole fund. Another parameter that may influence the predescriptors pertinency is the quality of the UNIfs composing this predescriptor. The filtering1 operation can be manual or automatic.

[1] The lengths of these fragments are here left as open questions.

A second filtering is applied. The filtering 2 operation is an extraction with respect to a variable threshold ρ_i of a subset from the predescriptors set. In fact, for each couple (ρ_i, q_j) a subset of predescriptors is selected from the predescriptors set according to a certain criterion. Hence, a family of predescriptors are selected by the MFP machine. These selected predescriptors become descriptors.

The second type of operation touches upon the quering procedures that measure the distances between a query and a set of descriptors in a dynamic image. Although not all researchers of IR believes in the effectiveness of linguistic analysis [17], we think that some morpho-syntactic analysis is pertinent in a dynamical system of IR. The summarized approach that we present is based on a paper of Pêcheux [12]. Our aim is to evaluate the distance between a query and a descriptor considered both as two sentences[2] . The proximity measure between a query and a descriptor 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 is the reason 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.

Definition 3.2.1. The edition distance between a couple of words $(U,V) \in V^*$ (free monoid generated by the vocabulary set V) is defined as:

$$\gamma^*(U,V) = \min_{s \in \Delta} \gamma(s) \;,$$

where Δ is the set of elementary operation series which transform the word U into the word V.

The proximity measure γ^* is a distance in the topological sense. The similarity between sentences uses two different costs, the first cost is based on weights given to parts of speech assigned to words and the second is based on the words themselves composing the sentences.

Let $P = p_1/f_1\ p_2/f_2 \ldots p_k/f_k \ldots p_n/f_n$ and $Q = q_1/t_1\ q_2/t_2 \ldots q_k/t_k \ldots q_m/t_m$, be two series of canonical words with their parts of speech provided by a lemmatizer.

Definition 3.2.2. The distance between two sentences P and Q 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 [15]. An internal representation module can transform a query 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 descriptor. The same remark can be stated in the case of an elliptic situation.

---

[2] The choice of a sentence as a particular case is not a constraint inherent to the method, the reason is that a sentence can be considered as a semantically complete unity.

4. The neighborhood system and the query space

We introduce the image neighborhood system and the query space in this section.

4.1. The neighborhood systems between images

In a Markovian flow process, the configurations of the information in the documentary fund is seen as a set of images (dynamic images in our approach) forming a neighborhood based on a certain similarity between them. This similarity is measured through some type of proximity. In more formal terms, a neighborhood system can be defined in the following manner.

Definition 4.1.1. Let $\vartheta = \{I_1, I_2,..., I_k\}$ be a set of images (vertices), $V = \{V_I, I \in \vartheta\}$ is called a neighborhood system for ϑ if it is a subset of ϑ such that:

$I \notin V_I$

$I \in V_K \Leftrightarrow K \in V_I$.

The doublet (ϑ,V) is a hypergraph of order $k = \mathrm{card}(\vartheta)$ where a hyperedge is composed of all images which are neighbors according to some sense. In order to obtain a physical[3] interpretation of the Markovian Field structure, one has to give sense to the notion of neighborhood. We define the neighbors set of an image k for a fixed α ($\alpha > 0$) and a specific distance L between images as:

$V(\alpha)_{\text{image k}} = \{\text{image l} ; L(\text{image k}, \text{image l}) \leq \alpha\}$. The neighborhood system is based on a proximity measure between objects of the same type. That is to say, the proximity measure is applied with respect to the sets of descriptors contained in two neighbor images. The distance is formally defined in the following manner:

Definition 4.1.2. If I is the morpho-syntactic distance between two descriptors, then an α ($\alpha > 0$) level neighboring system between images can be given by:
$V_\alpha(\text{image u}) = \left\{\text{image v} ; \|\text{image u} - \text{image v}\|_M < \alpha\right\}$,where "M" stands for the "norm of the mean" and it is defined as:

$$\|\text{image u - image v}\| = \frac{1}{P_{\text{image u}} \cdot P_{\text{image v}}} \sum_i \sum_j I(\text{des i}, \text{des j}) \ ,$$

where the couple of ((descriptor i) , descriptor j)) belongs to the cartesian product ({image u} * {image v}). The values $P_{\text{image u}}$ and $P_{\text{image v}}$ are the global weights associated to each image. They can be written as:

$$P_{\text{image u}} = \sum_i P_{\text{descriptor i}} \cdot \chi_{\text{image u}}(\text{descriptor i}) \ ,$$

---

[3] The term "physical" must be understood as an interpretation which can be applied to images universe and conveys a sense in this universe.

where χ is the characteristic function defined as:

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

Definition 4.1.3. A subset $C \subseteq \vartheta$ is called a *clique* if every pair of distinct images in C are neighbors. With respect to graph theory, the clique order here is equal to 2.
An intuitive translation of a clique is a cloud of neighboring points each representing an image. Very often, when we want to identify some differences between groups, one expresses a degree of discrimination between these groups by using the variance metric. This is possible in an Euclidean space as it is the case in the projection of the descriptors space into the weights space. However, the descriptors space is not necessarily Euclidean. This is the reason why we adopt "the norm of the mean" between two images.

4.2. The query space

As outlined above, a query can be interpreted as a request to compare a hypothetical image with the set of images present in the documentary system. If so, one can apply a proximity measure between the query and the set of images. However, the nature of a query approaches the one of a descriptor but not the one of an image. Some metrics are basically applied to a descriptor and a query and then induced to the image. This is the reason why we define in the next section an other type of proximity measure between a query and an image. Summarizing the preceding notions and translating them in topological terms, wee can say that the clouds are made of images sharing some similarity between their respective descriptors. A query can be seen as an attempt to find to which cloud it would belong if it was projected in the images space, that is to say to which cloud could it be matched (see figure 2).

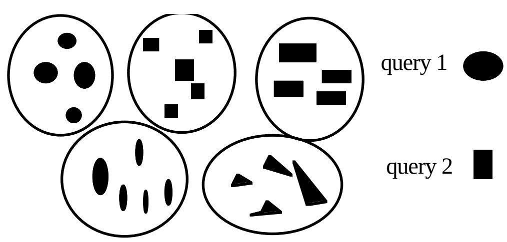

Fig. 2. Two queries have to be matched in the cliques space.

4.3. The query image matching

The matching of a query to the set of images can be defined in the following formal terms:

Definition 4.3.1. The degree of extracting images responding to a query execution is represented by a real function value called "the query image matching". The query image matching is taken as a proximity measure between the query and an image represented by a set of descriptors.

Definition 4.3.2. The matching function is a mapping M with the following analytical structure:

$$E * \{q\} \xrightarrow{M_q} IR^+$$

$$(des1, des2, \ldots, desN, q) \xrightarrow{M_q} \sum_{i=1}^{i=N} P_{des\ i} I(des\ i, q)\ ,$$

where $E \subseteq G$ is the set of descriptors, q is the query, $P_{des\ i}$ is a weight assigned to the descriptor i, N is the number of descriptors contained in an image.

4.4. A geometrical interpretation

Using the definition of the MRF, one can give a geometrical interpretation of an image in the space generated by the distance I. One can notice that the distance is similar to an L1 norm (sum of components). We can also use a norm similar to the L2 norm (quadratic) and in the case where the weights assigned to descriptors are all equal to P and X = R, R'=R/P one obtains a sphere whose radius is equal to R'. An arbitrary point A in the N-descriptors distance I space can be written as: $A = (I(des1,q),\ I(des2,q),...I(desN,q))^t$, where t is the transposed form. The distance space between descriptors and the query is modified in practise, because the weight P_i are different. One obtains ellipsoids or conics in general in the case of the quadratic MRF.

5. The Markov Random Field concept in IR

We present the query image MRF and justify the reason of using this concept.

5.1. The query image MRF

It seems much more consistent to study the query image matching not in all the documentary fund but only among a packet of images which are neighbors (according to the sense defined in section 4). For example, instead of covering all the images representing the documentary fund which is useless and much more expensive in cpu time, one has to cover only some particular set of neighboring images.

Definition 5.1.1. The query image matching M is a stochastic function, it is considered as a Markov Random field X with respect to V. It can be written:

$$X_{I_u}(\text{query}) = \sum_{i=1}^{i=N} P_{des\ i} I^2(des\ i, \text{query})\ .$$

where image $u = I_u$ is represented by the set {des1,des2 ,..., desN} of descriptors.

For easier mathematical interpretation, we use the quadratic distance to define this MRF. Let $X = \{X_I,\ I \in \vartheta\}$ denote any family of random variables indexed by ϑ. Let Ω be the set of all possible configurations: $\Omega = \{\omega = (x_{I1}, x_{I2}, ..., x_{Id})\}$, $I_i \in \vartheta$ and $i \in [1..d]$. If one abbreviates $\{X_{I1} = x_{I1}, ..., X_{Id} = x_{Id}\}$ as $\{X = \omega\}$ then it follows:

Definition 5.1.2. The variable X is *a Markov Random Field* with respect to a neighborhood system $\{\vartheta,V\}$ if:

Prob $\{X = \omega\} > 0$ for all $\omega \in \Omega$

Prob $\{X_I = x_I / X_r = x_r, r \neq I\}$ = Prob $\{X_I = x_I / X_r = x_r, r \in V_I\}$.

The definition of Markovian Fields by means of conditional probabilities has been first proposed by Dobrushin [5]. In this model, the MRF expresses the local interaction between images $\{I_j\}$ with respect to the query q. In other words, for a fixed query q, and a fixed image u (I_u), when one knows the matching degree of this query on all images of the fund except the image u, then the information about the matching degree of the image u with the query depends only on images which are neighbors to image i. Formally, this can be stated as:

$$\text{Prob}\,\{X_{I_u}(q) = x_{I_u} / X_{I_v}(q) = (x_{I_v})_{v\neq u}\} =$$

$$\text{Prob}\,\{X_{I_u}(q) = x_{I_u} / X_{I_w}(q) = (x_{I_w})_{I_w \in V_\alpha}\}\ .$$

One can notice that if we increase the values of I and P (weight) then the query image matching X will uncrease. Conversely, if we decrease the values of I and P then one will decrease the value of X. Hence, the MRF X appears like a *connection force* between a query and an image.

Theorem 5.1.3. *[Hammersley-Clifford] The random variable X is a Markov Random Field with respect to a neighborhood system $\{\vartheta,V\}$ if and only if Prob(ω) = Prob$\{X = \omega\}$ is a Gibbs distribution with respect to the same neighborhood $\{\vartheta,V\}$.*

Proof. See [9].

5.2. Why an MRF approach?

The hypotheses of using the concept of a Markovian Field in describing the group of similar images with respect to a flow of queries in a documentary fund originates from the inadequacy of classical methods [7] to take into account the dynamicity involved in the local interaction that goes on between images and queries. As outlined at the begining, in the classical model the relation in the set of images is stable and does not change with respect to the queries. The queries may change but the configuration of images in the documentary fund does not. The MFP hypotheses, on the contrary, allow to describe the influence a query may have on the configuration of the set of images. It does this by identifying how the neighborhood relations among images are influenced with respect to a flow of queries. In other terms, a Markovian Field allows each query to modify sighltly the regrouping of images. A series of queries may gather or separate the images differently. This latter move provides a change in the cliques (hyperedge) oganization. The configuration of groups of images are in constant change due to the introduction of the 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 [6] derived from energy which is easier.

6. The documentary system degradation

Our aim here is to provide a mathematical model capable to simulate the degradation of the documentary system with respect to the interrogator. This degradation depends at least on the threshold

ρ_i which in its turn depends on the flow of queries and the analytical structure of the morpho-syntactic distance I. The MFP machine aims to decrease this degradation by converging to lower energy states of the system configurations (as the majority of gradient dynamics). At this attractor state, modification of configurations becomes less and less possible. One parameter which may be responsible of this degradation is the distance I between a query and a descriptor. This distance can be corrupted by morphological and syntactic ambiguities assigned to descriptors and queries (different trees parsing, etc.)[2,3]. The distance I, the query q and ρ are involved in the definition of the Markov Random Field X. An other parameter which is inherent to the query nature is the degree of relevance of inquiries emanated from the user with respect to the domain encountered: this can be considered as: the competence level of the user with respect to the domain inquired and explored: it is the user degradation. At the first retrieval step, the documentary system interaction is considered as degraded. One can finally decompose this degradation into the following manner:

Total degradation = system degradation ⊗ user degradation,

where '⊗' is an invertible operator as '+' or 'x' etc.

Remark 6.1. Besides, the matching degree of images by a flow of queries is not necessarily uniform, it very often depends on the domain encountered.

This remark leads us to consider a weight function W which remains constant in a clique of images. Hence, for a fixed ρ_i and a fixed query q_j, one can express analytically the nonlinear degradation modeling of the representation of the fund as:

$$\mathrm{Deg}^{\rho_i}(q_j) = \psi\,(W(X^{\rho_i}(q_j))) \otimes N^{\rho_i}(q_j),$$

where $\psi(.)$ is a nonlinear invertible function as: (x->ln(x) or x---> sqrt(x)).

Analytically, this can be written as:

$$d^{\rho_k}_{q_i,I_j} = \psi\left(\sum_{I_v \in V_\alpha(I_j)} w^{\rho_k}_{q_i,I_v} \cdot x^{\rho_k}_{q_i,I_v}\right) + \sum_{I_v \in V_\alpha(I_j)} n^{\rho_k}_{q_i,I_v}\ ,$$

where w (weight) is constant in the set of images I_j contained in $V_\alpha(I_j)$. The random noise $n^{\rho k}{}_{qi,Iv}$ emanated from the user is supposed to be Gaussian with mean μ and variance σ^2. The matrix W= (w_{ij}) is defined as: $w_{\text{query } j\text{, image } u} = w_{\text{query } j\text{, image } v}$ if image v $\in V_\alpha$(image u).

7. The optimal configuration

Our aim is to determine an optimal configuration $\omega^* = (x_{I1}, x_{I2}, \ldots, x_{Id})$ and an optimal threshold $\rho^*$with respect to a flow of queries. Formally, the problem can be written:

$$\underset{\rho}{\mathrm{Max}}\left\{\underset{\omega}{\mathrm{Max}}\left\{\mathrm{Prob}\,(X^{\rho}=\omega\,/\,d^{\rho})\right\}\right\}$$

$$\Leftrightarrow\ \underset{\rho}{\mathrm{Min}}\left\{\underset{\omega}{\mathrm{Min}}\left\{U_{P}^{\rho}(\omega)\right\}\right\}\ ,$$

where "Up" stands for posterior energy which depends on the degradation d^{ρ}.

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. This probability is a Gibbs distribution and the problem is transformed into the research of a configuration ω and a threshold ρ at minimal energy states. By decreasing the temperature parameter T depending on two slow variables which are the threshold and the index query, we reach configuration states (fast variables) of lower energy. The temperature parameter T is a nonlinear function converging to zero when the number of queries increase and for a specific threshold ρ^*. In order to avoid local minima, we tolerate with a certain probability higher energy states: it is the simulated annealing algorithm [8] that we apply.

8. Conclusion

We have explored in this paper a Markovian Random Field approach as a possible alternative to the Information Retrieval problem. The MFP machine that we are developing is dynamic and adaptive to the user's queries. We notice that at each state of the machine, a set of cliques of images is formed by the interactivity proper to the machine. However, each clique has a different matching degree with a query and therefore the responses from the system to the user are not random. One for example can choose the maximum matching degree. As we outlined, the randomness is inherent to the nature of the query. Finally, we can outline that the classical precision and recall ratios must have an other theoretical interpretation in a dynamic machine.

References

[1] R. K. Belew, "Adaptive Information retrieval: Using a connectionist representation to retrieve and learn about documents", In N. J. Belkin (Ed.), Proc. 12th Annual International ACM SIGIR 1989.

[2] D. Bouchaffra, G. Lallich-Boidin and J. Rouault, "Echantillonnage stratifié avec restimation (Bootstrap): application à l'analyse morphologique", Secondes Journées Internationales d'Analyse Statistiques de Données Textuelles, Montpellier, 21-22 Octobre, 1993.

[3] D. Bouchaffra and J. Rouault, "Different way 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: AI & Statistics IV, Lecture Notes in Statistics Ser: Vol 89, Springer Verlag, 1994.

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

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

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

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

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

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

[10] 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.
[11] 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", in Revue de Liaison de la recherche en informatique cognitive des organisations (ICO Québec), Vol. 6 1 et 2,pp. 19-41, printemps 1993.
[12] M. Pêcheux, "Analyse automatique du discours", Dunod, 1969.

[13] V. Rijsbergen, "Information retrieval", 2nd. Edition, Butterworth & Cie. Publishers, Ltd., 1979.

[14] S. Robertson and P. Harding, "Probabilistic Automatic Indexing by learning from human indexers", J. Documentation, 40(4), pp 264-270, 1984.
[15] J. Rouault, "Analyse automatique du discours: Note sur le système 3AD75", Communication interne, Université Stendhal, Grenoble, 1994.
[16] G. Salton, J. Allan & C. Buckley, "Automatic structuring and retrieval of large texte file", Communications of the ACM, 37 (2), 97-107, 1994.
[17] G. Salton, C. Buckley and M. Smith, "On the application of Syntactic methodologies in automatic text analysis", Information Processing & Management, 26 (1), 73-92, 1990.
[18] M.William, "Connectionist models and Information Retrieval", 25, 209-259, 1990.