NOMAD: A Documentary Database Interrogation System Using Multiple Neural Topographies and Novelty Detection

Jean Charles Lamirel, Marion Crehange and Jacques Ducloy, CRIN-CNRS, Nancy, France

Abstract: NOMAD makes use of the synthesis capabilities and flexibility inherent in the connectionist approach to increase its processing power as compared to existing systems while proposing new operating modes directly accessible to a large number of users. NOMAD manages multiple synthetic type views on its documentary contents acting as case-memories as well as elaborated thematic browsing tools. NOMAD also manages a session memory with the following three functions: cumulative recording of user need, managing user contradictions and proposing new orientations. NOMAD has extended learning capabilities, enabling it to improve its performance in the long term.

1. Introduction

In the course of the interaction with an user, it generally seems difficult for current documentary systems to produce easily applicable synthetic views on the state of their knowledge, whatever the user profile. This remark also applies to the own system knowledge (thesaurus, thematic network), and to that acquired on the need of a given user in the course of his interaction with the system. Assimilating knowledge of thesaurus type indeed implies a long experience of utilization. Concerning thematic knowledge, which is not affected by the preceding defect, the current trend is to use it for browsing purposes rather than to integrate it in the general operation of the systems. The feedback is often a short-range process, reformulating queries on a case by case basis and thus cannot manage cumulative knowledge on a given need. The above mentioned difficulties are particularly disturbing for several reasons. A general reason is that the user often has a tendency to construct his need as his interaction with the system proceeds which makes it necessary to provide him with the greatest possible number of position marks to help him in his retrieval task. A more precise reason is that documentalists have shown that the user could perform retrievals of many types. In our opinion, some of the latter such as exploratory retrieval, themetic retrieval and connotative retrieval require the active use of synthetic knowledge, perhaps even the triggering of interactions between several types of knowledge synthesis in the course of exploratory and connotative retrievals.

On the other hand the use of a connectionist model for the design of a documentary database interrogation system has a certain number of important general advantages. Due to the fact that it allows the adoption of unique modes of reasoning and representation for the whole knowledge set processed by the system, a connectionist model considerably enhances synthesis capabilities, tolerance to noise and system evolution possibilities. These criteria seem to be essential in the documentary domain, where knowledge is naturally subjected to significant description noise and where user queries, which are the input data, are themselves approximate. It may be noted that connectionist models, related to the family of the so-called associative bidirectional models, have been used many times to reproduce under optimization the general operation of the conventional documentary systems with a vectorial approach (Biennier90). Nevertheless it is common to observe difficulties regarding focalization of the answers from these models, due to the weak granularity of the manipulated knowledge, not to mention the multiple propagation possibilities between the knowledge items. Connectionist models adapted to non-supervised learning, such as the Kohonen model, have been used for producing automatic contents analyses of documentary databases combined with document classifications. Axial component analysis (Lelu93) and topographical classification (Lin91) both make use of a Kohonen model associated with competitive learning. These two methods make it possible to produce synthetic knowledge maps which can be used as thematic browsing tools.
In our opinion, the prospect of giving to topographic classifications an active role during the interrogation seems to be full of promise. As a matter of fact, considering the topological properties (Kohonen84) associated with the latter, they may be regarded as true case-memories. However we have previously pointed out that elaborate thematic reasoning required an interaction process between multiple classifications. If topographical classifications are used as elementary classifications, it then becomes necessary to set up an interaction model between these classifications.

2. Genesis and functionalities of the NOMAD system

In our opinion, it is not possible to totally meet the previously defined user needs and at the same time to take advantage of the connectionist models, under direct use of existing models. We therefore propose a documentary database interrogation system based on an original neural model in which some components of existing models have, however, been integrated. Thus, the NOMAD system includes a general architecture model with neural components, a neural model of session memory comprising functionalities for managing contradictions, a neural model of multiple intercommunicating maps built on the basis of the elementary topographic map model which has been described above. While the general management of the interrogation session is maintained, the resulting system makes it possible for the user to combine numerous approaches which can be useful when formulating his query. In a classical way, the latter can use the formulation by descriptors, the selection of documents representing examples or counter-examples of its need and hypertext browsing between documents; he can also widen its point of view by proceeding in a similar way with the thematic knowledge associated with multiple topographic classifications. The user will then carry out selection of theme and counter-theme as well as inter-thematic browsing between different semantic domains. On the other hand, the NOMAD user can at any time have access to and modify the contents of a session memory which, by means of an analysis of his choices and rejections, provides him with a synthetic description per semantic domain of his need including its new possible orientations.

![NOMAD Architecture](image_url)
The NOMAD system design is based on an object-oriented approach, allowing a clear distinction between the roles allotted to its components. It takes advantage of standardization both in the representation strategy for the basic knowledge and in the user interaction:

The thematic knowledge and the session memory are similarly partitioned into **semantic domains**. A semantic domain can be defined as a consistent point of view of the knowledge to be described. A set of domains could thus be generated from the partition (strict or loose) of the totality of descriptor terms of the documents of a given universe (for example, it is possible to make the distinction between the terms describing the physical characteristics of an animal from those describing its mental characteristics in the universe of documents describing animals). Another set could be generated from the evaluation of profiles describing direct relationships between documents (animals have profiles of friend animals, enemy animals). The concept of semantic domain has led to the definition of two operating modes in the NOMAD system: In the **general mode** documents and queries are handled considering all their descriptor point of view in the same way. In this mode the user may give an opinion on a document, his decision will then be fed back in an identical way in the totality of the active semantic domains. In the **partitioned mode** documents and queries are handled by considering separately each of their respective descriptor points of view. This mode makes it possible for the user to define his need in a more accurate (also less noisy) way, thus improving the reasoning accuracy of the system.

In partitioned mode, a user can for example give the opinion that the description of a given animal meets his need only from the point of view of the physical characteristics.

Although reasoning is made on a continuum, all user decisions (about themes covered by the classifications, as well as about documents or descriptors) are discretized in five categories: **full acceptance (FA), moderated acceptance (MA), no opinion (NO), moderated rejection (MR), full rejection (FR)**. Discretizing the decisions has the advantage of being a natural approach for the user, for whom it is often difficult to think "analogically". It also opens the way to the use of symbolico-connexionist methods for the user need and the user behaviour management. Such a method has been chosen for the management of user contradictions in NOMAD.

In the following description of the **NOMAD components** we more particularly dwell on the role descriptions of the session memory manager and of the thematic knowledge manager.

The main role of the **session manager** is to control the progression of an interrogation session. This role consists in selecting and managing the general interrogation strategies of the system. (The interrogation strategies of the system may vary during a session according to user behaviour and the current results of the session. As an example, repeated contradictory behaviour of the user when formulating his query should lead to a preponderance of the thematic interrogation which is then directed by the classifications). It also consists in checking the respective states of the various system components (query manager, thematic knowledge manager, session memory manager) and managing communication between these components. The session manager equally manages long term learning at the end of the session according to the achieved results. The **query manager** role mainly consists in centralizing the information on the current query (global state of query, thematic knowledge and the session memory) and the user decisions at a given step of the session and in supervising the elementary construction of each new query during reformulation process. The **document manager** role is to supervise access to the expanded description of the documents and to their contents. It also supervises the hypertext browsing between documents.

The main role of the **thematic knowledge manager** is to control thematic reasoning of the system and user access to thematic knowledge. Thematic knowledge is represented by a set of classifications. Each classification of the set, which is associated with a semantic field, is represented by a Kohonen bidimensional topographic map. Thematic reasoning takes place in many cases. Among other:

a) **query thematic completing**, which consists in highlighting the 'best' themes associated with active semantic domains which were not present in an original query (If the user query
contains only the physical characteristics of an animal, the system will discover all the animal synthetic life styles which match at best, relative to the database content, with the user physical characteristics description).

b) **query general thematic recentering**, which consist in finding the ‘best’ themes connected to an original query in all the semantic domains (If the user query is concerning large predators, the system can discover, thanks to the classifications, only the presence of middle size predators and huge predators in the database. It will then propose these themes as possible synthetical adaptations for the query).

c) **retrieval oriented by the existing themes**, which is different from the preceding one in that the existing themes represent a full retrieval constraint. It can be initiated after a document browsing and selecting phase carried out by the user or initiated by the system itself after detection of repeated contradictory behaviour of the user.

d) **invitation to increase the number of documents** produced in answer to a query, under consideration of the profile of the themes which are in the neighbourhood of the ‘best’ themes of the maps.

All the results of the preceding operations are directly visible by the user on the maps. They can be submitted to his critical examination or directly managed by the system. Each topographic map is initially built up by unsupervised learning carried out on the whole documentary fund which is managed by the system. This learning takes place through the profile vectors extracted from the documents, which describe the characteristics of these documents in the map associated semantic domain. For each neuron of the map, the basic competitive learning fonction has the form:

\[
W_{t+1}^k = W_t^k + a(t) \times (p^t_d - W_t^k)
\]

where \(W_t^k\) is the external weights profile of the neuron \(n_k\) at time \(t\),

\(p^t_d\) is the description, in the semantic domain \(s\) of the map, of the document \(d\) chosen as learning sample at time \(t\),

\(a(t) = \frac{k_1}{t} \times e^{-k_2 \times \text{Edist}(n_k, n_\star)\}}\)

with \(k_1, k_2\) fixed positive constants,

\(\text{Edist}(n_k, n_\star)\) the Euclidean distance on the map between the current neuron \(n_k\) and the winning neuron \(n_\star\) at time \(t\), \(n_\star\) being such that \(\min_{n_{\text{Map}}} (p^t_d - W_t^n)\).

After the preliminary learning phase, each map is organized so as to be legible for the user: the map is divided into 'topologically consistent' zones through analysis of the main components of the neuron profiles. Each zone, which can be regarded as a macro-theme of synthesis (each neuron on the map being itself a theme of synthesis) with an associated name, is managed by an zone neuron. The topological properties associated with the Kohonen maps make it possible to project the original documents onto a map so that their proximity on the map matches as closely as possible their proximity in the semantic domain described by said map (two physically close animals are close on the map representing a view of physical characteristics). Once built, each map provides a point of view which is both a semantic domain specific and a synthetic view of the documentary fund that it covers. Given that it preserves an active neural structure, a map can play a double role. For the user, it can act as a direct navigation support in a given fund. For the system it can act as a computing support for the query thematic reformulation carried out on this documentary fund. Communication between maps is organized via document neurons. It takes place during the mutual interaction process between maps which is initiated by the thematic knowledge manager for thematic reasoning. The purpose of the mutual interaction process, which is based on a Grossberg-like adaptative resonance model, is to generate a new set of active zones. Various parameters associated with the neurons, such as the connection depletion speed or the activation remanence of the neurons, make it possible to monitor the type of effect (reformulation, completion) linked with this process.
The original documentary database contains descriptions of 50 dogs. The user's initial query only contains the descriptor 'hunting' pertaining to the semantic domain 'function'. The system discovered, among other things, that a high intelligence (int. ++ ) as related to mental characteristics is an important synthetic characteristic of the hunting function. This information can then be used in several ways. Moreover the topography directly highlights a strong correlation between high intelligence and low aggressivity (agr -). Neurons (themes) are represented by circles, their activity being simulated by a grey scale level. They have an associated number which represents their associated document count. The consistent zones are delimited by the border lines on the map.

The \textit{session memory manager} processes each user decision concerning themes, documents and descriptors in order to:

a) \textit{Highlight preferential directions} useful for describing the user need.

b) \textit{Highlight optimal optional directions}. Optional directions, i.e. directions that are not explicitly formulated in the user need, are useful both for the user and for the system. They can especially be helpfull for the latter during query reformulation if the use of preferential directions has not led to sufficient results.

c) \textit{Evaluate the user contradiction and precision rates}. The user feedback process (selection of examples and counter-examples) naturally leads to an increased noise and/or to a reduced accurracy in the system description of the user need. Nevertheless if the accuracy becomes to low and/or the noise becomes to high, these phenomena can surely be imputed to the user, who has then not been able to correctly formulate his need. In most existing systems this fact in not taken into account despite of its great importance both for correcting queries, for choosing a strategy which is well-suited to the experience level of each
individual user and for maintaining the consistency of the system knowledge when its evolution in the long term depends on the results of the interaction with said user.

The session memory manager stores its information in a session memory, which has been partitioned into semantic domains according to the same pattern as for the maps. The session memory of the NOMAD system is based on the novelty detector model derived from the works of Kohonen (Kohonen84). A novelty detector can be regarded as a neural filter which, after learning on a reference data set, acts as a projection operator in the vectorial subspace orthogonal with the vectorial subspace which supports the reference data set. For any new input data, the filter output value is called the novelty component of the data. This component is equal to zero if the data belongs to the vectorial subspace which supports the reference data. The component which is orthogonal to the novelty component is known as the “habituation” component. Each session memory includes two detector types: rejection detectors and acceptance detectors. Rejection detectors use as reference samples the themes, documents, and descriptors that the user presents as counterexamples of its need in the semantic domain. According to their general behaviour, such detectors give preference to new directions, i.e., those that are not present in their reference samples. As compared with the rejection detectors the acceptance detectors have a dual behaviour. Their reference samples are the themes, documents, and descriptors which have been presented by the user as examples of his need. Their behaviour is to give preference to the so-called “habituation” directions which are present in their reference samples. Semantic domain memories play an important part during the query reformulation. They make use of the results of the filtering, through their detectors, of the query descriptors associated with their domain for generating a new descriptors set (see following formulas). They communicate with maps via document neurons (reference documents) associated with their detectors. Decisions taken by the user on the documents, as soon as they are validated, modify the polarity and the weights of the connections between the documents and the maps.

A novelty detection type learning on a set of reference samples \( \{x_1, \ldots, x_i\} \), in which each \( x_i \) is a document or theme description vector on a semantic domain \( D \), or also a descriptor unit vector on \( D \), produces a filter \( \phi \) which converges to \( \phi^* = I - X_i^T X_k \), where \( X_k \) is the sample matrix, \( X_i^* \) is the Penrose pseudo-inverse of \( X_i \). \( \forall v \) defined on \( D \), \( \phi \times v \) is the \( v \) component \( \bot \) to the subspace of \( D \) spanned by the reference samples. The reformulation of a request \( R \) on \( D \), thanks to the detectors states, can be expressed as:

\[
R^o = R_0^o + \alpha_{FA} \times (I - \phi_{FA}) \times R_0^o + \alpha_{MA} \times (I - \phi_{MA}) \times R_0^o + \alpha_{MR} \times \phi_{MR} \times R_0^o + \alpha_{FR} \times \phi_{FR} \times R_0^o
\]

where one filter (detector) associated with a positive coefficient \( \alpha \) is tied to each decision category and \( \alpha_{FA}, \alpha_{MA}, \alpha_{MR}, \alpha_{FR} \).

A memory trace parameter is associated with each detector. This parameter defines the maximal number of reference samples that can be processed by a detector before eliminating the oldest samples from its learning phase. Generally speaking, the use of detectors with short-time memory will stress the mobility of the user decisions, while detectors with long-time memory will stress the integrity checking of the user decisions. The flexibility of the session memory model of NOMAD has also made it possible to set up strategies with various levels of generality, in order to solve contradictions which may arise from the user decisions. These strategies are based on analyses of the competitive activation and inhibition effects produced by the detectors on descriptors of their respective semantic domain. A user view on a semantic domain memory has the form of a synthesis of the state of its associated detectors which is elaborated as a partition of the domain descriptors between the novelty subspace, the “habituation” subspace, which are both associated with the above mentioned detectors, and a neutral state. Thus, the user can make use of the novelty, “habituation” and neutrality informations to modify its query with respect to the specific semantic domain.
Figure 3: View on the session memory associated to the semantic domain of dogs mental characteristics after several document selections.

The user has selected several examples and one counter-example of hunting dogs in the 50 dogs database. The memory state highlights that the user performed contradictory decisions as related to the domain of the mental characteristics of its query because the 'high intelligence' (intelligence++) descriptor was learnt both as an habituation descriptor for the full acceptance as well as for the full rejection decision types. The contradiction management mechanism will then be launched in the concerned domain. It will result in eliminating the 'full rejection' sample (caniche) of the domain memory because of the strong correlation existing between 'hunting' function, which is present in all the 'full acceptance' samples, and high intelligence (see figure 2). Each one of the four columns of the memory represents a detector associated to a decision type.

Long term learning is performed at the end of each interrogation session. A principal learning aimed at modifying the weights of the connections between the documents and the position of the latter on the maps. This learning is based on a calculation of the activity correlation between documents derived from the general results of the session (general state of the session, maps and detector state in each semantic domain) and on a contents analysis of the activity memories of the document neurons and of the zone neurons. A secondary learning aimed at weighting the influences of the semantic domains, the influences of the detectors in each domain, the sensitivities of the maps associated with the domains, the sensitivities of zones on each map associated with a domain and lastly, the interactions between these zones. In order to improve system knowledge in a consistent way the system takes into account the contradiction rate of the user during the session. A high contradiction rate will reduce the extent of the learning.
3. Prototype

A prototype of the NOMAD system has been realized in the SMALLTALK object-oriented programming language on a SUN Sparcstation as part of the PhD thesis of the first author. This prototype is currently being tested on two different types of documentary database. The first test database contains a thousand indexed images, representing old photographs of Paris, coming from the French Ministry of the Culture. This basis provides a good reference since it has already been used for the development and testing of the RIVAGE system, developed by our team. The second database is the the INRIA bibliographic database, with 25 000 references in TEXTO format. In this case, for the sake of performance, the learning process is only conducted on a typical subset of the database, whereas the access to the whole database content is achieved through an on-line communication process with an external document management toolkit. For this toolkit, we have chosen the DILIB SGML based toolkit that we developed for the first time in INIST (Ducloy91), and that we are further developing in eRIN. The DILIB toolkit, whose kernel consists in a SGML tree handling library, allows the creation of basic Information Retrieval Systems. It also includes various format conversion functionalities. As the basic data organization (where performance is paramount) is generated by the toolkit, we have only had to rewrite in the SMALLTALK programming language the access functions to SGML records for interfacing the SMALLTALK applications with whole database information.

4. Current developments and perspectives

As this domain is full of promise, we are continuing to thoroughly study the operating and learning strategies of our system. As the dimensions of the topographies processed by our system are not limited, we are planning to make use of rather strongly multidimensional topographies to achieve the two following targets. Firstly, to make it possible for the system to work on more precise classifications, each system classification having its own bidimensional equivalence in the universe of the user. Secondly, to represent the strategic knowledge of the system (user profiles, query typologies, ...). This approach is connected with the expansion of the model towards the generalized case based reasoning approach. Finally we are planning to use a dynamic thesaurus whose role will be both to assist the user in the formulation of his query and to make it possible for the system to improve its performance through managing the neural connections between descriptors. The use of such a thesaurus combined with the one of novelty detectors in a symbolico-connectionist way makes it possible for us to implement sophisticated learning methods. We will then be able to detect on a abstract level the user contradictions and the user need and so to better adapt the system strategy to the experience and to the need generality levels of the respective users.

References

Biennier F. (1990). Modélisation d'une base d'hyperdocuments et méthode connexionniste d'aide à la navigation. Thèse de doctorat de l'INSAL, Lyon, France


