A Type 2 Fuzzy Set Based Model for Adaptive Information Retrieval

Giacomo Piccinelli, Marco Casassa Mont
Internet Business Management Department
HP Laboratories Bristol

HPL-98-27
February, 1998

E-mail: [giapicc, mcm]@hplb.hpl.hp.com

fuzzy set, information retrieval, adaptivity, DBMS

The abstraction model that an information retrieval (IR) system imposes on data determines its knowledge on them. Uncertainty permeates the behaviour of both the system and the users and we argue that its explicit modelling is fundamental in order to increase classical IR parameters like precision and recall as well as to reduce the gap between systems and users.

We present a keyword based model that, capitalising on the flexibility of fuzzy sets, extends the traditional two dimensional vector approach to data abstraction, evolving it into a paradigm where relevance is tightly coupled with uncertainty and the view the system has on data evolves dynamically through an adaptivity process. A prototype system (DUNE) has been derived from the general model and we are currently evaluating possible applications of our proposal in the knowledge management field.

A Type 2 Fuzzy Set Based Model for Adaptive Information Retrieval

Giacomo Piccinelli and Marco Casassa Mont

Extended Enterprise Laboratory
Internet Business Management Department
Hewlett-Packard Laboratories
Bristol BS12 6QZ, U.K.

Email: {giapicc, mcm}@hplb.hpl.hp.com

Abstract

The abstraction model that an information retrieval (IR) system imposes on data determines its knowledge on them. Uncertainty permeates the behaviour of both the system and the users and we argue that its explicit modelling is fundamental in order to increase classical IR parameters like precision and recall as well as to reduce the gap between system and users.

We present a keyword based model that, capitalising on the flexibility of fuzzy sets, extends the traditional two dimensional vector approach to data abstraction evolving it into a paradigm where relevance is tightly coupled with uncertainty and the view the system has on data evolves dynamically through an adaptivity process. A prototype system (DUNE) has been derived from the general model and we are currently evaluating possible applications of our proposal in the knowledge management field.

1 Introduction

Managing huge amounts of information is crucial for modern information systems. The volume of data available in electronic format increases constantly and row data need to be turned into information to become process drivers. Information retrieval (IR) tools are the very base for any process that deals with big data bases, even when they support only data collection and the actual task of extracting the information is left to the user (human being or software agent).

The infrastructure data are plunged into has a dramatic impact on the effectiveness of an IR system and the internal data abstraction model it enforces is a key aspect. When data are structured, and the kind of abstraction we are interested in is close to their original structure (ex. invoices retrieved by progressive number), traditional data base management systems (DBMSs) enforcing relational or object oriented models offer an effective support. Forms map quite easily into objects or tables but a more complex kind of abstraction is required in order to enforce convenient views and access mechanisms for other types of data.

Expressiveness and adaptivity are fundamental features for a data model. The abstraction associated with an object should capture all its peculiarities in an easily manageable representation but deciding which are the "relevant" features of the object is difficult. The way in which an object is perceived by an observer depends on his/her interests and capabilities and they may evolve quite rapidly. An abstraction paradigm should allow different views on an object and, at the same time, it should support their refinement and evolution.

The aim of our work is to tackle both uncertainty and adaptivity problems through an integrated theoretical infrastructure based on fuzzy sets [13, 15]. After an overview of the main concepts underneath fuzzy set theory, we present a model for dynamic abstractions, based on “type 2” fuzzy sets, enforcing the dynamic link between relevance and uncertainty in keyword based description systems. We also present DUNE, a prototype based on our model, and we briefly discuss potentialities and open issues related to our proposal.

2 Elements of fuzzy set theory

The binary paradigm allows the direct modelling of a great number of problems and in many situations we can transform a problem in a binary equivalent with acceptable loss. The concept of binary choice is at the very base of many theoretical frameworks, ranging from set theory to predicates logic, but there are situations in which we need to consider a range of choices wider then "true" or "false" for an accurate modelling of the problem. Thinking of common sets, an element can either belong to a set or not and all the elements have the same belonging degree. Set theory describes a set as the collection of all the elements for which a given (binary) predicate holds true and this definition actually deals with a world of elements that is split in two by an ideal line: we distinguish an element only from the side in which it lays. We can consider a number of predicates at the same time and look at the intersection area but this solution becomes quickly unmanageable when the number of predicates grows. What we would like to do is to take our world of elements and to associate an element coming from a potentially different world to each of them depending on some sort of criteria: we can now partition our elements looking at their associated element. Without losing in generality, we can associate to each element a real number in the range [0,1] and the association law may be easily extended to cope with the change: we can imagine some sort of “level lines” linking the points with the same associated element. The step back to normal sets is simple: we only need to restrict ourselves to {0,1} as associated world.

Definition: *Given a pair of standard sets B and M, a* ***fuzzy set*** *F based on B is a pair (B, f) where f: B →M.*

In the usual terminology [15]: *B* is the "base set" or "support", *M* is the "membership space" and *f* is the "membership function" mapping any element of the support in the correspondent membership value. When M is the interval [0,1] the fuzzy set is normalised. The membership function is the main component in the definition: intersection, union, complement, cardinality as well as other concepts of standard set theory are transferred to fuzzy sets working on f [15].

The basic definition of fuzzy set [13] can be generalised and the simplest extension is the recursive use of fuzzy sets in the definition of the membership space. The concept of type is introduced for a fuzzy set in order to express the "depth" of its membership space [15].

Definition*: Let us consider a normalised fuzzy set as having "type 1". A "***type m***" fuzzy set is a fuzzy set with base B whose membership values are type m-1 (m>1) fuzzy sets with base [0,1].*

Thinking at the membership value associated to an element of the support as a description for that element [6], “type m” fuzzy sets introduce an hierarchical structure on the description where each component at one level may be further specified in the lower levels: the deeper the hierarchy, the more precise the description. The definition of basic operations, as well as metrics, needs to be adapted to the peculiarities of this kind of extension but this discussion is outside the scope of our work [15].

For our purposes, we are mainly interested in type 2 fuzzy sets on top of which we will define problem specific metrics and operations.

3 Addressed problem: Uncertainty and Adaptivity

Given a set of data (data base) and an information need (expressed through a query), the basic functionality of an IR system is to select the pieces of data out of the data base that may be useful in order to satisfy the information need. The problem maps into is the retrieval of data whose abstraction matches the description of an ideal object inferred from the query [2].

Solutions based on flat sets of keywords [8, 9] are widely used but they have intrinsic limitations due the impossibility to express different degrees of connection between the keywords and the object they represent. Information related to uncertainty and relevance are not reflected in the view the system has of the row data. Extensions of the keyword model can be found in the work of Salton [4, 12] and later developments [3, 4, 5] on vector representations where the idea of “weight” is introduced for the strength of the relationship between keyword and object.

Weights usually are modelled with real numbers and what happens is that semantically different information are compressed in a single number. We can suppose, for example, to use weights in the range {1,2, .., 10} and to deal with the book “Egyptian Secrets”. Saying that we are 100% sure that the keyword “water” has relevance 3 is different from saying that we are 30% sure that is has relevance 10. In a standard solution it is very likely that “water” is associated to the book with weight 3 in both cases. Moreover, we can have two groups of observers and one of them may suggest the weight for water is 1, because they are interested in agriculture and the book gives a tourist description of the rivers, while the other group may suggests a weight 10, because they are tourists and there are nice pictures of the rivers. We do not want water to receive a weight around 5 because when people from the first group look for water resources documents they may retrieve our book, and they do not want to, while it is difficult that tourists interested in nice places near some river can find it. This kind of problems deteriorates both precision and recall of an IR system and is explicitly managed in our model.

Another characteristic of an IR system that is usually underestimated is its ability to improve its knowledge keeping it updated. In a number of situations, namely the interaction with users and other agents, the system has the opportunity to receive some feedback and a careful use of this resource is the key for a slow but continuous evolution. As an example, confidence about the relevance of a keyword in a description should increase if the keyword repeatedly prove to be relevant while we should reduce the confidence value for a keyword when it doesn't prove to be of interest. Again, it is important to keep different views apart in order to avoid that, back to the first example, the comments of 1000 satisfied tourists affect the research activity of agriculture experts.

Adaptivity, together with uncertainty management, should be at the very base of an information retrieval system: the integration of these two aspects is the main line of our model.

4 Model specification

In the previous sections, we have pointed out how uncertainty modelling and adaptivity are fundamental aspects of an information retrieval system, especially when it has to manage a huge object base preserving precision and completeness. We present an integrated approach to both uncertainty and adaptivity problems based on type 2 fuzzy sets and the result is a model in which both static and dynamic aspects coexist and support each other. Keywords are still at the base of the abstraction model but, together with relevance information, we enrich them with information on confidence degree and we plunge the result into a dynamic management system. We first present the static information model and then the evolution mechanisms.

4.1 Object Description

Given an object O, our purpose is to obtain a compact but comprehensive description O_d of it. A limitation of classic vector representation is that it doesn't recognise the importance of uncertainty as a crucial component of the information we model. Assigning a small relevance value we model the fact the word could be eliminated from the description without informative loss: this may be because it actually doesn't describe the object, because we are not sure about it or a mixture of the two. What actually happens is that a single numeric parameter collects the information on both the relevance of the word in the description and the confidence we have on the correctness of our relevance estimation. There are situations in which both this parameters need to be considered at the same time but it is in general preferable to keep them apart and to merge them with specific procedures on a case by case base. In our model we propose something more because we enforce an estimation of the confidence on every possible relevance value for a word.

What we propose is a fuzzy set based construction that models, in a single point close to the keyword, different views on its relevance and the correspondent reliability. This gives us a dramatic advantage for the definition of similarity concepts between two descriptions [7, 10] but it proves to be useful also for the dynamic aspects of the model.

Definition: *An **object description** O_d is a pair composed of a type 2 fuzzy set FS (W, f) and a value ε we call experience. W is a set of keywords and the function **f** maps every $w \in W$ in a fuzzy set RFS ([0,1], ρ) where $\rho : [0,1] \rightarrow [0,1]$.*

The RFS fuzzy sets is a relevance descriptor that represents what can be seen as a "confidence distribution" over the normalised set of relevance values: each word has its own descriptor. We can assume that the absence of a word from W is equivalent to its presence in association to an RFS where ρ is a constant function that returns the smallest real number greater than 0. The fact that ρ has value 0 in an interval [x, y] means that its behaviour in [x, y] is unspecified. We can also assume f extended over any super-set Ω of W where it returns a dummy RFS for every w in Ω - W. The experience parameter ε is fundamental for the adaptivity features of the model as it gives an indication on the “strength” of the present status of the description. When we collect some feedback on O_d suggesting to change (to adapt) part of it, we can refer to ε in order to establish the scope of the change.

From the definition of O_d, we notice that the emphasis is on the ρ functions. They collect the actual information on the confidence distribution and they represent the crucial point to work on for both retrieval and adaptivity processes. We give no general specifications on their structure but we suggest that probabilistic tools can be used in the initial phase (see the prototype) of the O_d history while statistic tools are more appropriate during its remaining lifetime.

If we think of every O_d as a point in a complex descriptions space, we need to impose some sort of metric on that space in order to manage concepts like similarity between descriptions that are fundamental in the perspective of clustering and retrieval activities. For this purpose we introduce a binary function D_σ (we call it "distance function") that compares on a component by component base two object descriptions summarising the result in a numeric value.

Definition: *Given a pair of functions ρ_1 and ρ_2 where $\rho_i : [0,1] \rightarrow [0,1]$ for $i \in \{1,2\}$ and $\{[x^i, y^i]\}_{i=1..n}$ the set of intervals in [0,1] where the value of both ρ_1 and ρ_2 is not 0, we define the support function d_σ as*

$$d_\sigma(\rho_1,\rho_2) = \sum_{i=1..n} \int_{[xi,\, yi]} (\rho_1(x) - \rho_2(x))^{2\sigma}\, dx$$

Given a pair of object descriptions $O_d1=(W_1, f_1)$ and $O_d2=(W_2, f_2)$, we define the ***distance function*** *D_σ as*

$$D_\sigma\,(O_d1, O_d2) = \sum_{w \in W1 \cup W2} d_\sigma(f_1(w), f_2(w))$$

The function exploits all the knowledge on relevance and associated confidence accumulated in the fuzzy set structure in order to take into consideration all the views on every component of the description. For the unspecified parts one of the ρ functions we assume a perfect matching with the other one. The σ parameter is a positive integer value that decides the sensitivity of the function: the bigger it is, the lower is the amplification of the differences between the components.

The definition of σ is an important step for the system and the trade-off is between precision and recall: a small value for σ results in high precision, because even small differences are relevant, but, for the same reason, it negatively affects the recall.

4.2 Adaptivity

Given an object, characteristics that are of interest now may change in the future and the same may happen to the relations among the objects. The continuous evolution of the abstraction layer allows the system both to keep the pace with the user needs and to increase the lifetime of the data. We introduce adaptivity at the very bases of an IR system.

The solution we enforce takes advantage of the object description structure (O_d) and the interaction with the environment. The evolution process is based on abstraction comparison. If for the same object we have an O_d (S) from the system and an O_d (U) from the user, the idea is for the system to learn from the user. This doesn't mean that the system has to accept completely the user point of view replacing S with U but that we need to find an appropriate balance. In general, we need a sort of "unification" mechanism that merges two O_d in a meaningful way: the solution we propose is to link the weight of an O_d to the experience ε and to compute a weighted average value for all the components.

Definition: *Given two object descriptions O_d1 $<\varepsilon_1, (W_1, f_1)>$ and O_d2 $<\varepsilon_2, (W_2, f_2)>$ we define $M_{\alpha\,\beta}$ (O_d x O_d → O_d) the* ***merging function*** *in α and β (real functions) as follows:*

$$M_{\alpha\,\beta}(O_d1, O_d2) = < \varepsilon, (W, f)>$$

where

$$\varepsilon = \alpha(\varepsilon_1) \blacklozenge \beta(\varepsilon_2) \qquad W = W_1 \cup W_2$$

and, for all w in W:

$$f(w) = \mathrm{RFS}\ ([0,1], \rho)$$

where, given

$$f_1(w) = ([0,1], \rho_1) \quad \text{and} \quad f_2(w) = ([0,1], \rho_2)$$

we have

$$\rho = \frac{\alpha(\varepsilon_1)\ \rho_1 + \beta(\varepsilon_2) \cdot \rho_2}{\alpha(\varepsilon_1) + \beta(\varepsilon_2)}$$

This process merges the knowledge coming from different points in a unique O_d structure. A major problem is how to minimise the information loss while paying more attention to the information that is, in some sense, more valuable (more reliable). The experience value ε is a good reference for the maturity of the information coded into an O_d but a number of external elements may affect the evaluation process. Therefore, we introduced the adjustment parameters α and β (the process is not guaranteed to be symmetric). For the binary operator ♦ we have a range of choices depending on the policies we enforce: simple solutions are +, *max* or *min*. We can use these parameters in order to enforce ageing policies, security policies or source selection policies and, in this sense, we suggest the possibility to

take advantage of user profiling, per user or per class of users, for a comprehensive plan on the α, β to use in different situations.
The position of the merging function within the model becomes clearer looking at its applications and the more important is in combination with the distance function for the management of clusters and adaptivity. Every time we are able to associate one (or more) objects to a description we need to test, using the distance function, if we have similar views: if this is the case, we invoke the adaptive association procedure.

Definition: *Given two object descriptions O_dS and O_dU for the object O, the adjustment functions α (for O_dS) and β (for O_dU) and two real values* min *and* max*, considering $\delta = D_\sigma$ (O_d S, O_dU) we define the* ***adaptive association*** *process as follows:*

- *if the δ is less than the threshold* min*, we associate the object O to O_dS*

- *if the δ is greater than* min *but smaller than* max*, we can merge O_dS and O_dU using the merging function M with parameters α and β and we associate the object O to the result of the merge*

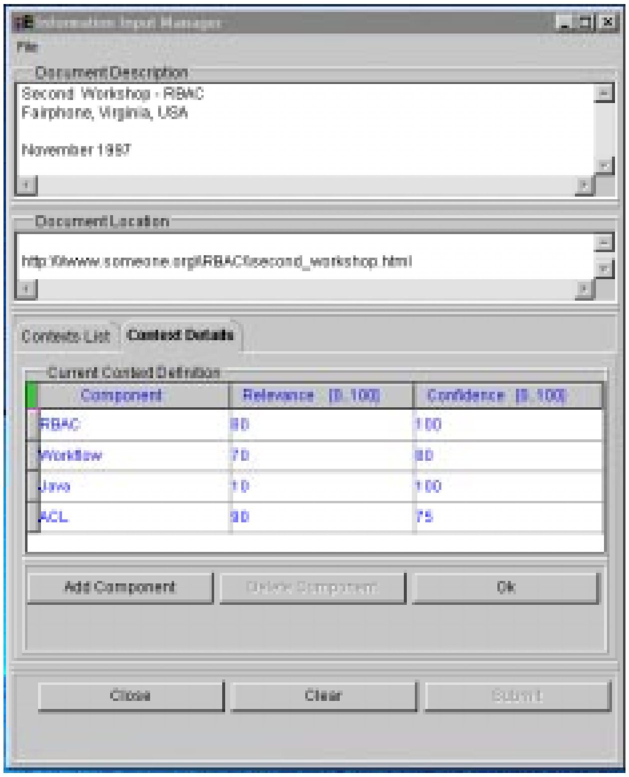
- *if the δ is greater than* max*, we associate the object O to both O_dS and O_dU*

Again, *min* and *max* are fundamental parameters as they affect space and time complexity together with the system precision and recall. The impact on clustering depends on the absolute values for min and max while adaptivity aspects are more related to the gap between min and max.
The proposed solution may be further refined but simplicity has to be kept as a guide in any choice: the enforcement of a continuous evolution process requires the steps to be simple in order not to reduce the overall performance of the system.

5 Prototype

In this section we briefly describe the prototype of an IR system we built in order to investigate the impact of our proposals on a real architecture. DUNE (Description UNcertainty and Evolution system) deals with an object base made of documents and it covers both the abstraction and retrieval aspect of data management. The choice of the document (a possibly structured piece of text) as basic data type is for generality reasons but the core mechanisms are independent from the nature of the objects assumed that they can be described by a set of attributes.
The metaphor we use for the interaction with the user is “context” based because it allows the users to focus on relevance and confidence aspects of the keyword avoiding, for the moment, more sophisticated compositional mechanisms. Although the details behind the user interface are outside the scope of this work [11], we enforce the idea that information on relevance and confidence for the keywords has to be as direct as possible in order to avoid complex and error prone heuristics.
In our view, a context is a set of keywords in association with an indication of their relevance and the confidence on such relevance: the transposition of the metaphor into the interface is immediate (Fig.1, 2). Looking at the actual interaction between the system and the outside world, we have two typical situations: when a new document is added to the object base (Fig.1) and the retrieval of the documents related to a particular problem (Fig.2). In the first case, the user is requested to

specify a number of contexts the document may be related to while, in the second case, the user describes its information need through a context and the system will retrieve all the matching documents.



Document Description

Second Workshop - RBAC
Fairphone, Virginia, USA

November 1997

Document Location

http://www.someone.org/RBAC/second_workshop.html

Current Context Definition

Component	Relevance [0..100]	Confidence [0..100]
RBAC	80	100
Workflow	70	80
Java	10	100
ACL	90	75

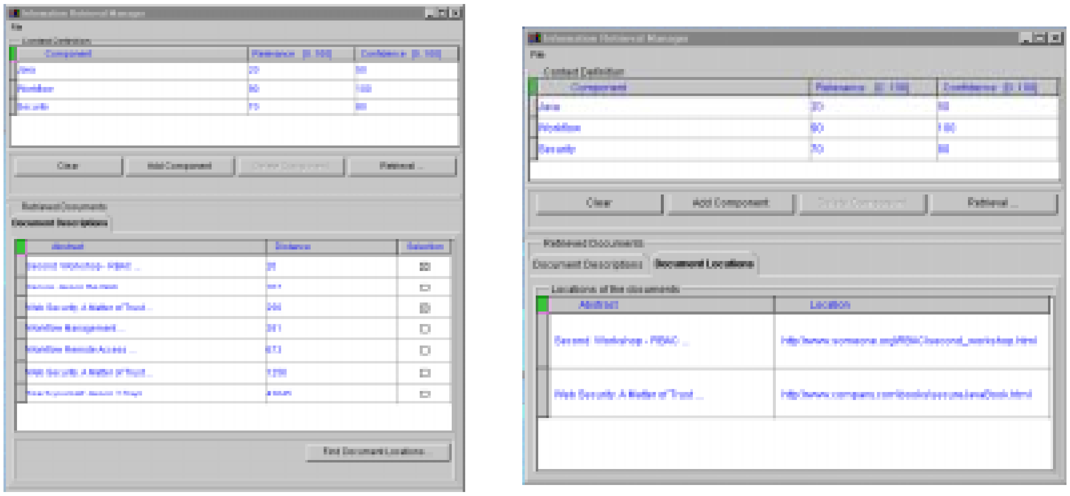
Fig.1: Document abstraction interface.

Concerning the feedback from the users on the adequacy of the answer to the query, the system infers it from the documents they actually selects (Fig.2). The users are first presented with a brief abstract of the document, they select those they are interested in and eventually they obtain the actual data. From the fact that the user selects a document associated to a specific object description O_s, we infer that the description O_q of the ideal object derived from the query can give us some information on how to evolve O_s that will undergo the adaptivity process. The core of the system implements the solutions proposed in the previous sections. Concerning the choice for the ρ function in the RFSs, we tried to balance expressiveness and simplicity so we decided for a Gauss-bell shaped function of the form:

$$\rho\,(x) = e^{-k(x-a)^2}$$

where *"a"* is the more significant relevance value and *"k"* expresses the confidence we have on relevance values different from *a* (*e* is the Napier number). This confidence becomes smaller as we move away from *a* but *k* decides the reduction speed. In Fig.3 we have a graphical representation of ρ for different values of *a* and *k* where we can see the bells shrinking as *k* grows: thinking of the meaning of ρ we understand how this influences the view on an object.

The fact that in this solution ρ is mono-modal (in the general case ρ is multiple-modal) introduces some limitations in terms of expressiveness but we are well compensated in terms of efficiency and this positively affects the scalability of the system. Tested on diagnostic data, particularly sensitive to uncertainty, the full version of the prototype registered improvements up to 30% compared with a version implementing the basic vector model.



Information Retrieval Manager

File

Context Definition

Component	Relevance [0..100]	Confidence [0..100]
Java	20	50
Workflow	80	100
Security	70	80

Abstract	Distance	Selection
Second Workshop - RBAC ...	35	☒
[illegible]	111	□
Web Security: A Matter of Trust ...	249	☒
Workflow Management ...	311	□
Workflow Remote Access ...	673	□
Web Security: A Matter of Trust ...	1250	□
[illegible]	[illegible]	□

Information Retrieval Manager

File

Context Definition

Component	Relevance [0..100]	Confidence [0..100]
Java	20	50
Workflow	80	100
Security	70	80

Abstract	Location
Second Workshop - RBAC ...	http://www.someone.org/RBAC/second_workshop.html
Web Security: A Matter of Trust ...	http://www.company.com/books/secure/JavaBook.html

Fig.2: Query interfaces (requests and result)

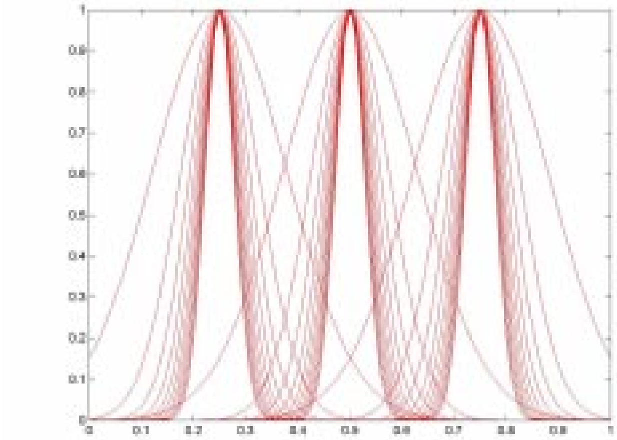


Fig.3: Graphic representation for ρ functions

6 Future Works

Bearing in mind the adaptivity principle, the uncertainty information gathered in the object description may be exploited in many ways and at different abstraction layers. Although we focused primarily on low level parameters like precision and recall, the final goal is to turn data into knowledge and it will be interesting to see how uncertainty based information bases could support classic knowledge management systems [14]. Concerning implementation aspects, efficiently manageable multiple-modal functions and patterns for optimised data access structures present interesting field of investigation.

6 Conclusions

Uncertainty is present in many aspects of information retrieval and the process of extracting from the object a meaningful and efficiently manageable abstraction is perhaps the more sensitive. Expressiveness is fundamental for any abstraction model especially when it is the only bridge between object base and problems space: the more information we have on the objects, the more accurate are retrieval and management processes. In this perspective, the model we propose extends the classical "vector model" along two directions: adaptivity and reflexivity. A lot of elements may affect the computation of the relevance of a keyword for the description of an object and our model enforces the possibility to express them through a complete relevance distribution instead of irreversibly collapsing all the information on relevance and confidence on a single value. In this way, it is also possible to express different views on the same component of a description and this information may be exploited in the retrieval process, that is still based on the matching of object descriptions (system knowledge) and an ideal description derived from the user query.

Adaptivity is an important aspect of the model and a constant evolution of the system view on the data base is enforced through a continuous meta-data evolution and acquisition process. Concerning the prototype (DUNE), it is based on a simple instance of the model and an essential interface but it nevertheless gave us interesting feedback on the results of explicit uncertainty management coupled with evolutionary techniques.

Bibliography

[1] A. Bookstein. Fuzzy requests: an approach to weighted Boolean searches. In *Journal of the American Society for Information Science*, Vol. 31, 1981.

[2] R. Bellman, R. Kalaba and L. Zadeh. Abstraction and pattern classification. In *Fuzzy Models for Pattern Recognition*, IEEE Press, 1992.

[3] C. Buckley and N. Fuhr. Probabilistic document indexing from relevance feedback data. In *Proc. 13th Int. Conf. ACM SIGIR on Research and Development in Information Retrieval*, Bruxelles, 1990.

[4] C. Buckley, G. Salton and G.T. Yu. An evaluation of term dependence models in information retrieval. In *Research and Development in Information Retrieval (LNCS 146)*, Berlin, 1982

[5] D.A. Buell and D.H. Kraft. Performance evaluation in a fuzzy retrieval system. In *Proc. of the 4th Int.Conf. on Information Retrieval*, Berkley, California, 1981.

[6] A. Kaufmann. Introduction to the theory of fuzzy subsets. Academic Press, 1975

[7] D.H. Kraft and D.A. Buell. Fuzzy set and generalised Boolean retrieval systems. In *Readings in fuzzy sets for intelligent systems*. Edited by D. Dubious,

H. Prade and R.R. Yager, 1993

[8] C.D. Paice. The automatic generation and evaluation of back-of-book indexes. In *Prospects for Intelligent Retrieval*, Informatics 10, 1989.

[9] S.E. Robertson and S. Walker. On relevance weights with little relevance information. In *Proc. 20th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Philadelphia, 1997.

[10] T. Radecki. Fuzzy set theoretical approach to document retrieval. In *Information Processing and Management*, Vol. 15, Pergamon Press, 1979.

[11] R. Rao and alt. The information grid: a framework for information retrieval and retrieval-centered applications. In *Proc. 5th Symposium on User Interface Software and Tech. (ACM UIST)*, Monterey. 1992.

[12] G. Salton and R.K. Waldstein. Term relevance weights in on-line information retrieval. In Information Processing and Management, Vol. 14, Pergamon Press, 1978.

[13] L.A. Zadeh. Fuzzy sets. In *Information Control*, Vol. 8, Academic Press, 1965.

[14] L.A. Zadeh. The role of fuzzy logic in the management of uncertainty in expert systems. In *Fuzzy Set and Systems*, Vol. 11, North Holland, 1983.

[15] H.J. Zimmerman. Fuzzy set theory and its applications (2° Edition). Kluwer Academic Publishers, 1991.