Fuzzy-Set Based Information Retrieval for Advanced Help Desk

Giacomo Piccinelli, Marco Casassa Mont
Internet Business Management Department
HP Laboratories Bristol
HPL-98-65
April, 1998

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

Help desk,
fuzzy sets,
information retrieval,
customer support

The effectiveness of a help desk system strongly depends on the ability to handle huge amounts of information. The precision of the information retrieval (IR) phase has a direct impact on both length and quality of the solution process and the improvement of precision is the focus of our research.

Capitalizing on "type 2" fuzzy sets, we propose a keyword-based IR framework where relevance and confidence are explicitly modeled both in the system-user dialog and in the knowledge base management. Adaptivity features assure a constant evolution of the system knowledge and experiments prove that the proposed techniques actually improve the precision of the IR process.

Fuzzy-Set Based Information Retrieval for Advanced Help Desk

Giacomo Piccinelli and Marco Casassa Mont

Hewlett-Packard Laboratories
Bristol BS12 6QZ, U.K.

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

Abstract

The effectiveness of a help desk system strongly depends on the ability to handle huge amounts of information. The precision of the information retrieval (IR) phase has a direct impact on both length and quality of the solution process and the improvement of precision is the focus of our research.

Capitalizing on "type 2" fuzzy sets, we propose a keyword-based IR framework where relevance and confidence are explicitly modeled both in the system-user dialog and in the knowledge base management. Adaptivity features assure a constant evolution of the system knowledge and experiments prove that the proposed techniques actually improve the precision of the IR process.

1 Introduction

Help desk systems are designed to provide customer support through a range of different technologies [19] and Information Retrieval (IR) tools play a fundamental role in this activity. Efficiency and effectiveness in data retrieval are crucial for the overall problem solution process but they depend on the infrastructure data are stored into and the correspondent abstraction model. The abstraction associated with an object should capture all its peculiarities into an easily manageable representation but deciding which are the "relevant" features of an object is complex and uncertainty makes this task even harder.

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 help desk systems and fuzzy sets, we present a model based on “type 2” fuzzy sets enforcing the dynamic link between relevance and uncertainty in keyword based Information Retrieval systems. Experimental results are presented and discussed.

2 Help Desk Systems

In a competitive business environment, customer satisfaction is a vital objective for many companies: high-quality products and high-quality customer service are absolutely two strategic aspects [18]. Help desk systems play an important role in this context, providing customer support and functions [19] like change, configuration and asset management (Fig. 1). The two main components of a help desk system are the front-end and the back-end ones: the former manages the interaction with customers while the latter deals with information retrieval issues. Usually help desk front-ends are implemented accordingly to two basic architecture [16]: single point of contact (SPOC) or multiple point of contact (MPOC) (Fig. 1). Both of them are deeply dependent on the back-end components of a help desk: the Information Retrieval system, the Knowledge Base and the set of procedures used to store information (Fig.1).

Focusing on Information Retrieval system, implementation issues are critical both for the overall performance of the system and the accuracy of the retrieved information. Customers usually provide data with different degrees of confidence depending on how that information has been collected. Current Information Retrieval tools do not explicitly model the uncertainty associated to information but they “mix” the measure of relevance associated to information with the relative measure of confidence. They don't even manage the feedback provided by users about the accuracy and usefulness of the retrieved solutions.

The explicit management of relevance and confidence on information, integrated with an adaptivity process is the key factor to improve the retrieval precision of a help desk system. Our proposal addresses the above aspects using fuzzy sets as the enabling technique to achieve the goal.

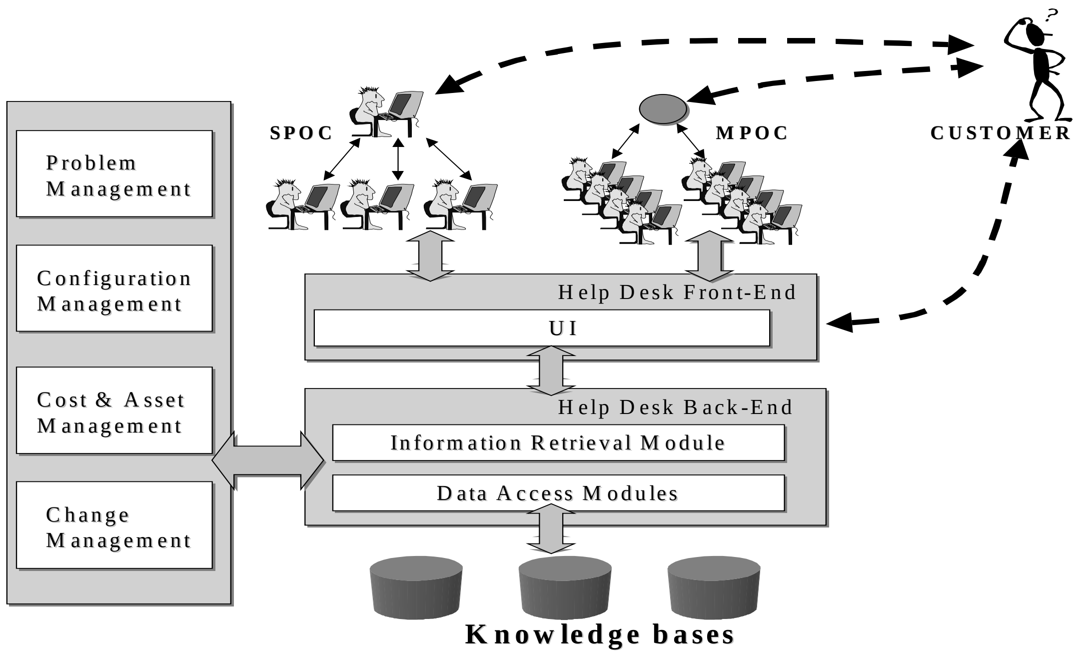


Fig.1: Help Desk components

3 Elements of fuzzy set theory

Problems usually can be modelled using the binary paradigm. However there are cases where we need to consider a broader set of choices instead of the only possible two: "true" or "false". Set theory describes a set as the collection of all the elements for which a given (binary) predicate holds true. Extending the above concept, we can associate each element of our world to another element belonging to a potentially different world: we can now partition our elements looking at their associated element.

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.*

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 [15]. It is possible to generalise the basic definition of fuzzy set [13] with a recursive use of fuzzy sets in the definition of the 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]*.

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

4 Uncertainty and Adaptivity management

Looking at IR system, the core functionality is the retrieval of data from a data base whose abstraction matches the description of an ideal object, inferred from a query [2]. Flat sets of keywords [8,9] are commonly used but they have intrinsic limitations. Extensions are proposed by Salton [4, 12] and later developments [3, 4, 5] where the idea of "weight" (usually modelled as a real number) is introduced for the strength of the relationship between keyword and object. That abstraction model shows its limits when it compresses semantically different information (like relevance and confidence) in a single number: both precision and recall are badly affected. Another limit of the abstraction model is in the ability to dynamically reflect feedback coming from user of the system: an effective use of that information is the key to enable a process of system adaptation.

We present an integrated approach to both uncertainty and adaptivity problems based on type 2 fuzzy sets. 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.

Given an object O, our purpose is to obtain a compact but comprehensive description O_d of it. Our proposal consists of a fuzzy set based construction that models and associates to keyword different views of its relevance and its correspondent reliability. This gives us an advantage for the definition of similarity concepts between two descriptions [7, 10] and 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* ρ: $[0,1] \rightarrow [0,1]$.

In order to manage concepts like similarity between descriptions, in the perspective of clustering and retrieval activities, we introduce a binary function D_σ ("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* ρ_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*

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

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*

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

We also 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. 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* $M_{\alpha,\beta}$ $(O_d \times O_d \rightarrow O_d)$ *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:

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

we have:

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

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*.

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

5 Prototype

The prototype focuses on an object base of documents describing solutions to customer problems. A document "context" is a set of keywords associated with their relevance and the confidence on such relevance.

We decided to implement the ρ function in the RFSs as a discrete function: given a keyword *k* in a context *x*, our ρ function models the confidence *c* of *k* over its relevance space. The multiple-modal nature of ρ allows us to capture different view of the confidence on *k*, reflecting its evolution as well. In (Fig. 2) it is depicted a possible evolution of the function ρ for a keyword *k* in a context *c*. The x-axis represents the value of relevance about *k* in a range [0,1] while the y-axis represents the correspondent value of confidence, in a range [0,1].

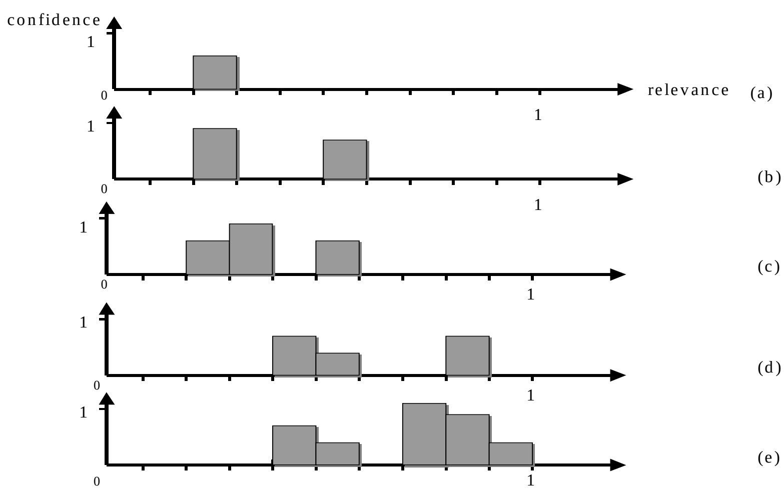
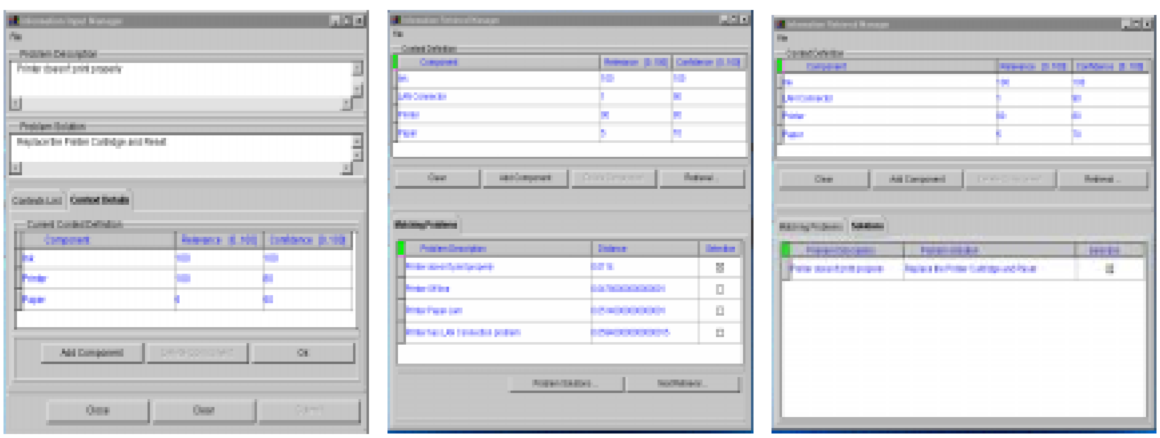


Fig.2: evolution of confidence over the relevance of a keyword *k*

The evolution from the state (a) to the state (e) (Fig. 2) shows that the relevance on the keyword *k* starts with a value of 0.3, with confidence 0.6, and it ends up with a strong confidence of the fact that the relevance of *k* is between 0.8 and 1. The sensitivity of the function can be tuned by modifying the number of steps in the relevance range.

Looking at the user interaction, two basic activities are allowed: adding documents to the object base and retrieving documents (Fig. 3). Users can add a solution to a problem (Fig. 3a) and the associated context reflects the description of the problem.

A user query is a set of keywords together with an explicit indication of their relevance and confidence (Fig. 3b). As a result, the user is presented with descriptions of “similar” problems. The user can chose the descriptions that better match its problem obtaining the related solutions (Fig. 3c). User selections also trigger the adaptivity process. The core of the system implements the solutions proposed in the previous section.



(a) (b) (c)

Fig.3: Document abstraction interface - Query interfaces (requests and result)

We tested the system under stress condition (up to 1000 solutions for a single problem) looking at clustering problems. Having fixed the solution set, we progressively increased the number of possible confidence levels from 1 to 5, looking at the average dimension of clusters and their overall number. We noticed that more than 5 different levels of confidence are of no practical use when an operator is a human being. The result of this experiment is shown in (Fig. 4).

In similar conditions, the number of solutions associated with a single problem is up to 5 times smaller than a system implementing standard IR techniques. This fact is reflected by the number of clusters and by their average dimension.

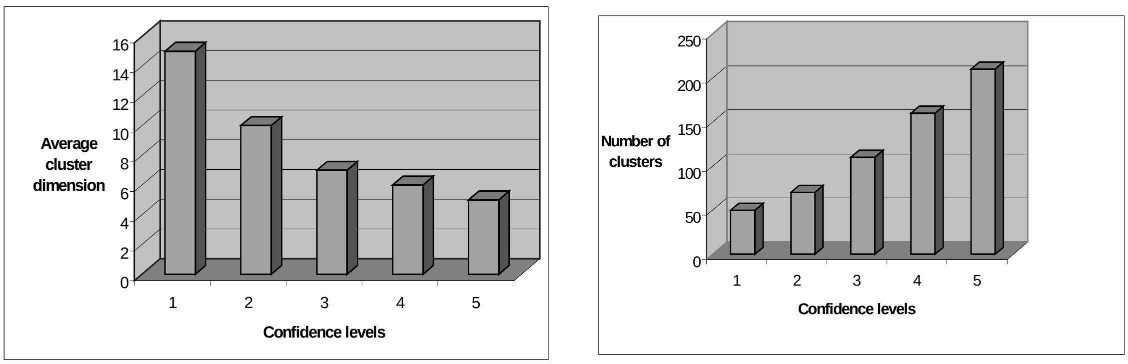


Fig.4: clustering

6 Conclusions

Information retrieval (IR) is a key aspect of help desk service. Uncertainty is a major factor in this context and the aim of our work is to provide a comprehensive solution for its management.

After a brief overview on help desk and the basic notion of fuzzy set, we present an IR framework, based on “type 2” fuzzy sets, in which relevance and confidence are used to enrich the descriptive power of keyword paradigm. We describe both the static and dynamic aspects of the model as well as a prototype in which all these components are turned into a fully featured IR tool.

Experimental results, especially in situations of "dense" data distributions, show that we obtain a clear improvement in terms of precision if compared to the results obtained with standard IR techniques in similar conditions.

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 modelsin information retrieval. In *Research and Development in InformationRetrieval (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.

[16] Ian Watson. Applying Case-Based Reasoning: Techniques for Enterprise Systems. CBR and Customer Service, pp. 89-115. Morgan Kaufmann Publisher, 1997.

[17] Maria LaTour Kadison, Blane Erwin, Michael Putnam. Internet Customer Service. Business Trade & Technology Strategies. Forrester Report, Volume One, Number Five, November 1997.

[18] Rita Marcella and Iain Middleton. The role of the help desk in the strategic management of information systems. OCLS Systems & Services, Volume 12 - Number 4 - pp. 4-19, MBC University Press, 1996.

[19] Frances Tishhler. The Evolution of the Help Desk. Telemarketing & Call Center Solutions Magazine, July 1996.

[20] TopicEDITOR Guide. Chapter 3 – Weights and Document Importance. Verity, Inc., 1996.

[21] Avron Barr. Knowledge Distribution at the Info Center Help Desk: The Future of Corporate Information Systems. Information Center Magazine, March 1991.