Mathware & Soft Computing 7 (2000), 1-15

Flexible information retrieval: some research trends

Gabriella Pasi
ITC-CNR, Via Ampère 56, 20131 Milano
gabriella.pasi@itim.mi.cnr.it

Abstract

In this paper some research trends in the field of Information Retrieval are presented. The focus is on the definition of flexible systems, i.e. systems that can represent and manage the vagueness and uncertainty which is characteristic of the process of information searching and retrieval. In this paper the application of soft computing techniques is considered, in particular fuzzy set theory.

Keywords: information retrieval, flexible query language, personalized document indexing.

1 Introduction

The huge quantity of multimedia information on the World Wide Web raises the need for efficient and effective systems that support an easy access to the information items relevant to specific users' needs. The activity aimed at locating on the WWW some relevant information is a very hard one: the users who access the network looking for something relevant to their needs can be seen as travellers opening a door on a wild forest, which has to be explored to the aim of reaching a (more or less) known destination. The users approaching the World Wide Web can access the big amount of available and mostly unknown information in different ways, which are also related to the purposes of their search. The most immediate approach is to directly navigate through the web sites by means of a chain of links found in the pages; in this way a formal expression of information needs is not necessary. However, when some specific information is searched, this point and click access paradigm is unpractical, and the effectiveness of the results strongly depends on the starting page.
The increasing efforts aimed at defining effective systems that help users to access information relevant to their needs witness the importance of this research field. The most known systems belonging to this category are Information Retrieval Systems

(on the web, the search engines) [13,26,30]. When using an IRS the users have to explicitly specify their needs by a formal expression of a query language through a set of constraints that the relevant information items must satisfy. The aim of an IRS is to estimate the relevance of information items on the basis of a comparison of the formal representation of the items with the formal user's query.

The activity of these systems is based then on the solution of a decision-making problem: how to identify the information items that correspond to the users' information preferences (i.e. *relevant* to their information needs)? What the users expect from an IRS is a list of the relevant information items ordered according to their preferences. The IRS acts then as an intermediary in this decision process: it "tries" to simulate the decision process that the user would personally undertake. To activate this automatic process the user has to formally communicate to the IRS her/his information needs, which are then analyzed by the system to the aim of selecting the information items satisfying them. The information items have to be formally represented to allow their automatic comparison with the formal user query. This representation problem is a very complex task pervaded by uncertainty and vagueness: the users' expression of information needs is often uncertain and vague, the formal representation of the documents' informative content introduces a loss of information and as a consequence is characterized by uncertainty about the real semantics of the documents' information content. The effectiveness of an IRS is therefore crucially related to the system's flexibility, intended as its capability to deal with the vagueness and uncertainty of the retrieval process, and to learn the user's concept of relevance through an adaptive behaviour. Commercially available IRSs generally ignore these aspects; they oversimplify both the representation of the documents' content and the user-system interaction.

Some important research efforts are aimed at defining systems tolerant to imprecision and uncertainty in the elicitation of users' preferences and able to learn them through an interactive and adaptive behaviour [1,4,6,10,11]. A big deal of research is being done in the area of intelligent information agents [23].

The aim of this paper is to synthetically present some approaches to the modeling of flexibility with respect to the previous mentioned aspects: the representation of information items and the definition of access method tolerant to vagueness and uncertainty in the specification of users' information needs.

In section 2 some approaches to model flexibility in IRSs are presented. In section 3 some applications of fuzzy set theory to model the vagueness and imprecision of the retrieval activity are synthesized. In section 4, a recent research approach to define personalized indexing mechanisms is analyzed. Finally, in section 5 some research approaches aimed at modelling flexible query languages are presented.

2 Modelling flexible Information Retrieval Systems

Information Retrieval (IR) aims at defining systems able to provide a fast and effective content-based access to a large amount of stored information [24,26,30].
Information can be of any kind: textual, visual, or auditory, although most actual IR systems (IRS) store and enable the retrieval of only textual information organized in documents. A user accesses the IRS by formulating a query, which the IRS evaluates to the aim of retrieving all documents which it estimates relevant to the query. The problem of identifying the information relevant to specific needs is a decision-making problem: when someone analyzes a huge amount of information items he/she has to decide which ones are relevant to his/her needs. The information items constitute the alternatives on which an evaluation process has to be performed to the aim of identifying the relevant ones [31].
An Information Retrieval System plays the role of an automatic intermediary in this decision process: its main objective is to estimate the items relevant to a specific user request: this requires a formal representation of both information items and user queries. The main components of these systems are: a collection of information items, a query language which allows the expression of selection criteria synthesizing the users' needs, and a matching mechanism which estimates the relevance of information items to queries (see Figure 1).
The input of these systems is constituted by a user query; their output is usually an ordered list of selected items, which have been estimated relevant to the information needs expressed in the user query.

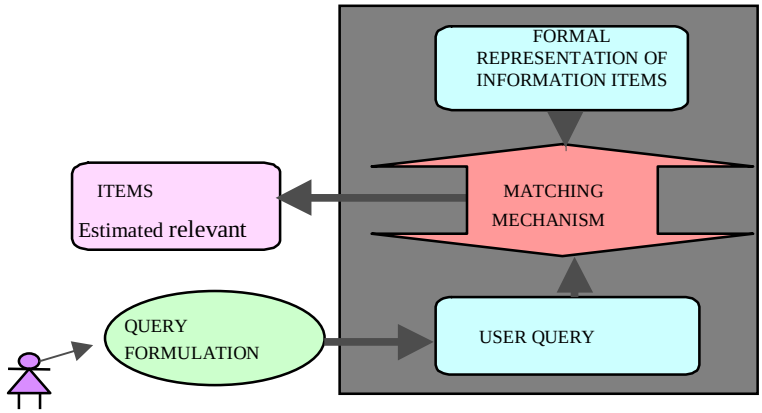

Figure 1: scheme of a system for the storage and retrieval of information

The ultimate aim of an IRS is then to estimate the relevance of documents to users' information needs. This is a very hard and complex task, since it is pervaded with imprecision and uncertainty.

Most of the existing IRSs offer a very simple modeling of IR, which privileges the efficiency at the expenses of the effectiveness. By formulating a query the user communicates to the system her/his preferences, on the basis of which an utility function can be defined and evaluated by the system (estimate of the alternatives' utility) [12]. A crucial aspect affecting the effectiveness of the system is related to the characteristics of the query language, which should represent in the more accurate and faithful way the user's information needs. The available query languages are based on keyword specifications, and do not allow to express uncertainty and vagueness in the specification of constraints that the relevant information items must satisfy [26,30].

Another important aspect which affects the effectiveness of IRSs is related to the way in which the information items are formally represented; the documents' representations are extremely simple, based on keywords extraction and weighting; moreover the IRSs generally produce a unique representation of documents for all users, not taking into account the each user looks at a document content in a personalized way, by emphasizing some subparts with respect to others. This adaptive view of the document is not modelled. Another important aspect is related to the fact that on the WWW some standard for the representation of semi-structured information are becoming more and more employed (such as XML); for this reason it is important to exploit their structure in order to represent the information they contain.

A promising direction to improve IRSs is to model the concept of partiality intrinsic in the IR process and to make the systems adaptive, i.e.able to "learn" the users' concept of relevance. In recent years big efforts have been devoted to the attempt to improve the performance of IR systems, and the research has explored many different directions to the aim of modelling the vagueness and uncertainty that invariably characterize the management of information [6,10,11].

A first research direction aims at defining methods of analysis of the natural language [28]. The main limitation of this approach is the level of deepness of the analysis of the language, and its consequent range of applicability: a satisfying interpretation of the documents' meaning needs a too large number of decision rules even in narrow application domains.

A second research direction is more general: its objective is to define retrieval models which deal with imprecision and uncertainty in the retrieval process. The most long standing set of approaches belonging to this class goes under the name of Probabilistic IR [9,30]. The aim of Probabilistic IR is to develop ad hoc models able to cope with the uncertainty of the retrieval process. However, there is another set of approaches receiving increasing interest that aim at applying techniques for dealing with vagueness and uncertainty. This set of approaches goes under the name of Soft Information Retrieval. The expression Soft Computing (SC) was introduced by Lotfi Zadeh as a synergy of methodologies useful to solve problems using some form of intelligence that divert from traditional computing. The principal constituents of SC are: fuzzy logic, neural networks, probabilistic reasoning, and evolutionary

computing, which in turn subsume belief networks, genetic algorithms, parts of learning theory, multivalued logics. As each of these methodologies allows to singularly representing imprecision, uncertainty and learning, it is frequently advantageous to employ these them in combination, rather than exclusively. SC differs from conventional (hard) computing in that, unlike hard computing, it is tolerant to imprecision, uncertainty, partial truth, and approximation. Because of these properties, SC can provide very powerful tools for IR. In [10] some techniques and applications of Soft Computing in Information Retrieval are presented.

Genetic Algorithms have been mainly applied to IR for improving document representation and indexing, and for defining relevance feedback mechanisms [10,15]. Evidential and Probabilistic Reasoning in IR has been mainly applied for defining IR models. There have also been some applications of techniques like Rough Set Theory and Multivalued Logics [11].

Neural networks have been used in the context of IR to design and implement IRSs that are able to adapt to the characteristics of the IR environment, and in particular to the user's interpretation of relevance [11,21]. In [11] some approaches concerning the application of connectionist approaches to the IR are analyzed. In particular, the two most important paradigms of learning used in the NN field are analyzed: the supervised learning and the unsupervised learning techniques. A supervised learning procedure is a process which incorporates an "external teacher". This means that the teacher specifies the desired output of the NN. During the learning phase the NN adapts the values of the weights on the connections in order to obtain the desired output [27]. In unsupervised learning procedures the NN does not receive any teaching or learning feedback, but it is left to learn by itself. This procedure is also often referred to as "self-organization" because the process relies only upon local information and internal control to learn by capturing regularities in the stream of input patterns. For these reasons, unsupervised learning has been used in IR mainly for documents or terms clustering and classification. In IR, documents or terms can be clustered in related groups so that, once identified a relevant one, retrieval of associated documents or terms can be facilitated. In [11] some approaches in this class of applications are reviewed.

Fuzzy set theory has been extensively applied to extend IR to model some aspects of the vagueness and subjectivity characterizing the retrieval process. In the next section the main applications of fuzzy set theory to IR are synthetically reviewed.

3 Fuzzy modelling of Information Retrieval

To the aim of defining flexible IRS, fuzzy set theory has been successfully employed to the following aims:

1. to deal with the imprecision and subjectivity that characterize the indexing process;

2. to manage the user's vagueness in query formulation;
3. to deal with discriminated answers reflecting the partial relevance of the documents with respect to queries;
4. to soften the associative mechanisms, such as thesauri and documents' clustering, which are often employed to extend the functionality of the basic IR scheme.

A survey of fuzzy extensions of IRSs and of fuzzy generalizations of the Boolean IR model can be found in [6,14].

Fuzzy generalizations of the Boolean model have been defined to the aim of defining IRSs able to produce discriminated answers in response to users' queries. In fact, Boolean IRSs apply an exact matching between a Boolean query and the representation of each document, defined as a set of index terms. They partition the archive of items into two sets: the relevant documents and the irrelevant ones. As a consequence of this crisp behaviour, they are liable to reject relevant items as a result of too restrictive queries, and to retrieve useless material in reply to general queries [26]. To the aim of softening the Boolean IR model, fuzzy set theory has been applied at distinct levels.

In documents' indexing some fuzzy techniques have been applied to the aim of providing more specific and personalized representations of documents' information content than those generated by the existing indexing procedures. In section 4, the fuzzy interpretation of the weighted document representation is introduced, and a fuzzy indexing model of documents structured in logical sections (such as XML documents) is presented. This model can be tuned by users on the basis of their personal criteria for interpreting the content of documents [4]. Also an indexing procedure for HTML documents is shortly described [19].

Fuzzy set theory has also been employed for defining flexible query languages, able to capture the vagueness of user needs as well as to simplify the user system interaction. This aim has been pursued at two levels: through the definition of soft selection criteria (soft constraints), which allow the specification of the distinct importance of the search terms. Query languages based on numeric query term weights with different semantics have been first proposed as an aid to define more expressive selection criteria [7,14]. Then, an evolution of these approaches has been defined, which introduces linguistic query weights, specified by fuzzy sets such as *important* or *very important*, in order to express the distinct importance of the query terms [2]. Another level of flexibility concerns the definition of soft aggregation operators of the selection criteria, characterized by a parametric behaviour which can be set between the two extremes AND and OR adopted in the Boolean language. In [3] the Boolean query language has been generalized by defining aggregation operators as linguistic quantifiers such *as at least k* or *most of*. These extensions are presented in section 5.

As it happens with search engines, the incorporation of a weighted document representation in a Boolean IRS is a sufficient condition to improve the system with

a document ranking ability. As a consequence of this extension the exact matching applied by a Boolean system can be softened to a partial matching mechanism, evaluating the degree of satisfaction of the user's query for each document retrieved. This value is called the Retrieval Status Value (RSV), and can be used for ranking documents. However, as it will be seen in section 4, more flexible indexing functions can remarkably improve the systems' effectiveness. The main idea is to explicitly model an indexing strategy that adapts the formal document representation to the user personalized view of documents' information contents.

Fuzzy "knowledge based" models [14,16], and fuzzy associative mechanisms based on thesauri or clustering techniques [14,17,18] have been defined in order to cope with the incompleteness characterizing either the representation of documents or the users' queries. In [17] a wide range of methods for generating fuzzy associative mechanisms is illustred. Fuzzy thesauri and pseudothesauri can be used to expand the set of index terms of documents with new terms by taking into account their varying significance in representing the topics dealt with in the documents; the degree of significance of the associated terms depends on the strength of the associations with the documents' descriptors. An alternative use of fuzzy thesauri and pseudothesauri is to expand each of the search terms in the query with associated terms, by taking into account their distinct importance in representing the concepts of interest; the varying importance is dependent on the associations' strength with the search terms.

Fuzzy clustering can be used to expand the set of the documents retrieved by a query with associated documents; their degrees of association with respect to the documents originally retrieved influence their Retrieval Status Value.

4 Personalized indexing in IR

The production of effective retrieval results depends on both subjective factors, such as the users' ability to express their information needs in a query, and the characteristics of the Information Retrieval System. A component of IRSs which plays a crucial role in determining their effectiveness is the indexing mechanism, which has the aim of generating a formal representation of the contents of the information items (documents' surrogates). The most used automatic indexing procedures are based on term extraction and weighting: the documents are represented by means of a collection of index terms with associated weights (the index term weights); an index term weight expresses the degree of significance of the index term as a descriptor of the document information content [25,26,29]. The vector space model, the probabilistic models and fuzzy models adopt a weighted document representation [26,30]. The automatic computation of the index term weights is based on the occurrences count of a term in the document and in the whole archive [26,30]. In this case the indexing function computes for each

document d and each term t a numeric value, by means of a function F; an example of definition of the function F is the following, in which the index term weight is proportional to the frequency of term t in the document d, and inversely proportional to the frequency of the term in the documents of the archive:

$$F(d,t) = tf_{dt} \times g(IDF_t) \quad (1)$$

where:

- tf_{dt} is a normalized term frequency which can be defined as: $tf_d = OCC_{dt}/MAXOCC_d$; OCC_{dt} the number of occurrences of t in d, and $MAXOCC_d$ is the number of occurrences of the most frequent term in d;
- IDF_t is an inverse document frequency which can be defined as: $IDF_t = log(N/NDOC_t)$, where N is the total number of documents in the archive and $NDOC_t$ is the number of documents indexed by t, g is a normalizing function. The computation of IDF_t is particularly costly in the case of large collections which are updated online.

The definition of such a function F is based on a quantitative analysis of the text which makes it possible to model the qualitative concept of significance of a term in describing the information carried by the text. The adoption of weighted indexes allows for an estimate of the relevance or of a probability of relevance of the documents to the considered query [26,30].

Based on such an indexing function and by maintaining the Boolean query language, the first fuzzy interpretation of an extended Boolean model has been to adopt a weighted document representation and to interpret it as a fuzzy set of terms [8]. From a mathematical point of view this is a quite natural extension: the concept of the significance of index terms in describing the information content of a document can then be naturally described by adopting the function F (such as the one defined in (1)) as the membership function of the fuzzy set representing a document. Formally, a document is represented as a fuzzy set of terms: $R_d = \Sigma_{t \in T} \mu_d(t) / t$ in which the membership function is defined as μ_d: $D \times T \rightarrow [0,1]$. In this case $\mu_d(t) = F(d,t)$, i.e. the membership value is obtained by the indexing function F. Through this extension of the document representation, the evaluation of a Boolean query produces a numeric estimate of the relevance of each document to the query, expressed by a numeric score, called the Retrieval Status Value (RSV), which is interpreted as the degree of satisfaction of the constraints expressed in a query.

The weighted representation of documents based on the F indexing function has the limitation of not taking into account that a term can play a different role within a text, according to the distribution of its occurrences. Let us think for example at scientific papers organised into the sections *title*, *authors*, *abstract*, *introduction*, *references*, etc. (this kind of structure can be explicitly defined by means of the XML language). An occurrence of a term in the *title* has a distinct informative role than an occurrence in the *references*. Moreover, indexing procedures based on the F function defined in (1) behave as a black box producing the same document

representation for all users; this enhances the system's efficiency but implies a severe loss of effectiveness. In fact, when examining a document structured in logical sections the users have their personal views of the document's information content; according to this view in the retrieval phase they would naturally privilege the search in some subparts of the documents' structure, depending on their preferences. This last consideration outlines the fact that relevance judgments should be driven by a user's interpretation of the document's structure, and supports the idea of *dynamic* and *adaptive* indexing [1,4,5]. By adaptive indexing we intend personalized indexing procedures which take into account the users' indications to *interpret* the document contents and to "build" their synthesis on the basis of this interpretation. It follows that if an archive of semi-structured documents is considered (e.g. XML documents), flexible indexing procedures should be defined by means of which the users are allowed to direct the indexing process by explicitly specifying some constraints on the document structure (preference elicitation on the structure of a document). This preference specification should be exploited by the matching mechanism to the aim of privileging the search within the most preferred sections of the document, according to the users' indications. The user/system interaction can then generates a personalized document representation, which is distinct for distinct users [1,4,5].

In [5] a user adaptive indexing model has been proposed, based on a weighted representation of semi-structured documents that can be tuned by users according to their search interests to generate their personal document representation in the retrieval phase. The considered documents may contain multimedia information with different structures. A document is represented as an entity composed of sections (such as *title*, *authors*, *introduction*, *references*, in the case of a scientific paper). The model is constituted by a static component and by an adaptive query-evaluation component; the static component provides an a priori computation of an index term weight for each logical section of the document. The formal representation of a document is a fuzzy binary relation defined on the cartesian product $T \times S$ (where T is the set of index terms and S is the set of identifiers of the documents' sections): with each pair <section, term>, a significance degree in [0,1] is computed, expressing the significance of the term in the document section.

The adaptive component is activated by the user in the phase of query formulation and provides an aggregation strategy of the n index term weights (where n is the number of sections) into an overall index term weight. The aggregation function is defined on the basis of a two level interaction between the system and the user. At the first level the user expresses preferences on the document sections, outlining those that the system should more heavily take into account in evaluating the relevance of a document to a user query. This user preference on the document structure is exploited to enhance the computation of index term weights: the importance of index terms is strictly related to the importance for the user of the logical sections in which they appear.

At the second level, the user can decide which aggregation function has to be applied for producing the overall significance degree (see figure 2). This is done by the specification of a linguistic quantifier such as *at least k* and *most* [33].

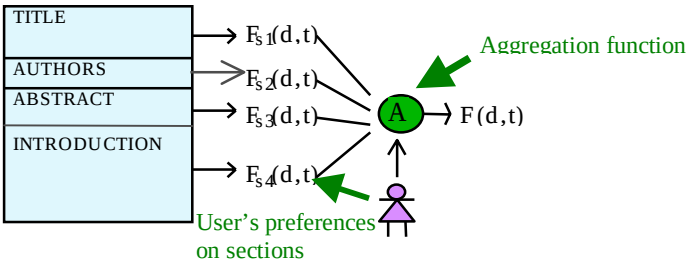

Figure 2: scheme of the personalized indexing procedure

By adopting this document representation the same query can select documents in different relevance orders depending on the user indications.
It is very important to notice that the elicitation of users' preferences on the structure of a document is a quite new and recent research approach, which can remarkably improve the effectiveness of IRSs.
In [19] another representation of structured documents is proposed, which produces a weighted representation of documents written in HyperText Markup Language. An HTML document has a syntactic structure, in which its subparts have a given format specified by the delimiting tags. In this context tags are seen as syntactic elements carrying an indication of the importance of the associated text: when writing a document in HTML, one associates a distinct importance with distinct documents' subparts, by delimiting them by means of appropriate tags. On the basis of these considerations, an indexing function has been proposed, which computes the significance of a term in a document by taking into account the distinct role of term occurences according to the importance of tags in which they appear.

5 Flexible query languages allowing soft constraints specification

A flexible query language is a language that makes possible a simple and approximate expression of subjective information needs. By means of fuzzy set theory some flexible query languages have been defined as generalizations of the Boolean query language. In this context a flexible query may consists of either both of the following two soft components or just one: the first component is constituted by selection conditions that are interpreted as soft constraints on the significance of the index terms in each document representation. The second component is

constituted by soft aggregation operators which can be applied to the soft constraints in order to define compound selection conditions. The atomic selection conditions are expressed by weighted terms expressed by pairs <term, weight>, in which weight can be either a numeric value in [0,1] (which identifies a soft constraint) or a linguistic value of the variable *Importance*, and the compound conditions are expressed by means of linguistic quantifiers used as aggregation operators. The notion of linguistic variable is suitable to represent and manage linguistic concepts and for this reasons it has been used to formalize the semantics of linguistic terms introduced in the Boolean query language to generalize it. When soft constraints are specified, the query evaluation mechanism is regarded as performing a fuzzy decision process that evaluates the degree of satisfaction of the query constraints by each document representation by applying a partial matching function. This degree (the Retrieval Status Value) is interpreted as the degree of relevance of the document to the query and is used to rank the documents. Then, as a result of a query evaluation, a fuzzy set of documents is retrieved in which the RSV is the membership value. The definition of the partial matching function is strictly dependent on the query language definition and specifically on the semantics of the soft constraints, and is defined as a bottom-up evaluation procedure: first, each atomic selection condition (soft constraint) in the query is evaluated for a given document, and then the aggregation operators are applied to the results starting from the inmost operator in the query to the outermost operator. The soft constraints are defined as fuzzy subsets of the set [0,1] of the index term weights; the membership value $\mu_{weight}(F(d,t))$ is the degree of satisfaction of the soft constraint associated with the query constraint denoted by *weight* by the index term weight. The result of the evaluation is a fuzzy set: $\Sigma_{d\in D}\mu_{weight}(F(d,t))/d$.

The first proposal to specify soft constraints was by means of numeric weights associated with terms. A numeric weight identifies a soft constraint on the weighted document representation, depending on the considered semantics. Distinct semantics have been associated with query weights [5,14]. However, the association of a numeric value forces the user to quantify the qualitative concept of importance of query weights, also if at the level of query evaluation this constraint is softly evaluated. To overcome this limitation and to make the query language more user friendly, in [2] a linguistic extension of the Boolean language is defined, based on the concept of linguistic variable [32]. By this language the user can associate with query terms either the primary term "*important*", or some compound terms, such as "*very important*" or "*fairly important"* to qualify the desired importance of the search terms in the query. When defining a query language based on linguistic query term weights, first the term set, i.e., the set of all the possibile linguistic values of the linguistic variable *Importance* must be defined: this definition depends on the desired granularity that one wants to achieve. The greater the number of the linguistic terms, the finer the granularity of the concepts that are dealt with. A pair <t, *important*>, expresses a soft constraint $\mu_{important}$ on the term significance values (the F(d,t) values). The evaluation of the relevance of a given document d to a query

consisting solely of the pair <t,*important*> is based on the evaluation of the degree of satisfaction of the associated soft constraint; this value is obtained by applying the function $\mu_{important}$ to the value F(d,t).

A second level of softening of the Boolean query language concerns the specification of aggregation operators. In the Boolean query language, the AND and OR connectives allow for crisp aggregations which do not capture any vagueness. For example, the AND used for aggregating M selection conditions does not tolerate the unsatisfaction of a single condition; this may cause the rejection of useful items. To face this problem, other extensions of Boolean queries have been provided, which concern the replacement of the AND and OR operators with soft operators for aggregating the selection criteria [20,26].

Within the framework of fuzzy set theory a generalization of the Boolean query language has been defined, based on the concept of linguistic quantifiers: they are employed to specify both crisp and vague aggregation criteria of the selection conditions [3]. New aggregation operators can be specified by linguistic expressions, with a self-expressive meaning such as *at least k* and *most of*. They are defined with a behavior between the two extremes corresponding to the AND and the OR connectives, which allow, respectively, requests for *all* and *at least one of* the selection conditions. The linguistic quantifiers used as aggregation operators, are defined by Ordered Weighted Averaging (OWA) operators [34].

By adopting linguistic quantifiers, the requirements of a complex Boolean query are more easily and intuitively formulated. For example when desiring that *at least 2* out of the three selection conditions "politics", "economy", "inflation" be satisfied, one should formulate the following Boolean query:

(politics AND economy) OR (politics AND inflation) OR(economy AND inflation)

which can be replaced by the simpler one:

at least 2(politics, economy, inflation)

The expression of any Boolean query is supported by the new language via the nesting of linguistic quantifiers. For example a query such as:

<image> AND (<processing> OR <analysis>) AND <digital>

can be translated into the following new formulation:

all (<image>, *at least 1 of* (<processing>, <analysis>), <digital>)

A quantified aggregation function can thus be applied not only to single selection conditions, but also to other quantified expressions.

In [5] a generalisation of the Boolean query language that allows to personalize the search in structured documents (as illustred in section 4) is proposed; both content-based selection constraint, and soft constraints on the document structure can be expressed. The atomic component of the query (basic selection criterion) is defined as follows:

$$aq = t \text{ in } Q \text{ preferred sections}$$

in which t is a search term expressing a content-based selection constraint, and Q is a linguistic quantifier such as *all*, *most*, or *at least k%*. Q expresses a part of the structure-based selection constraint. It is assumed that the quantification refers to the sections that are semantically meaningful to the user. Q is used to aggregate the significance degrees of t in the desired sections and then to compute the global Retrieval Status Value RSV(d,*aq*) of the document d with respect to the atomic query condition *aq*.

6. Conclusions

In this paper some approaches to the definition of flexible Information Retrieval Systems have been presented. In particular some promising research directions that could guarantee the development of more effective IRSs have been outlined. Among these, the research efforts aimed at defining new indexing techniques of semi-structured documents (such as XML documents) are very important: the possibility of creating in a user-driven way the documents' surrogates would ensure a modeling of the users' interests also at the indexing level (usually this is limited to the query formulation level).

The need for a flexible user-systems interaction characterizes also other systems which give a support to the access of relevant information, such as the systems which help users in online shopping. This kind of information access is mainly related to the definition of systems for the electronic commerce, and is aimed at selecting some products or services available on the network. When using these systems the main problem is that, usually, even if the customer's preferences are well defined, it is very difficult to select the possibly interesting products among thousands. The recommender systems are defined in order to propose to customers a selection of possibly relevant items, by providing them with a personalized access to information. In [22] the problem of modeling flexibility in this kind of systems is discussed.

References

[1] C. Berrut, Y. Chiaramella (1986). Indexing medical reports in a multimedia environment: the RIME experimental approach, ACM-SIGIR 89, Boston, USA, 187-197.

[2] G. Bordogna and G. Pasi (1993). A fuzzy linguistic approach generalizing Boolean information retrieval: a model and its evaluation, *Journal of the American Society for Information Science*, 44(2), 70-82.

[3] G. Bordogna, and G. Pasi (1995). Linguistic aggregation operators in fuzzy information retrieval. *International Journal of Intelligent systems*, 10(2), 233-248.

[4] G. Bordogna, G. Pasi (1995). Controlling retrieval trough a user-adaptive representation of documents. *International Journal of Approximate Reasoning*, 12, 317-339.

[5] G. Bordogna and G. Pasi (2000). Flexible Representation and Querying of Heterogeneous Structured Documents. *Kibernetika*, Vol. 36, N. 6, 617-633.

[6] G.Bordogna, G. Pasi, Modelling Vagueness in Information Retrieval (2001). In *Lectures in Information Retrieval*, M. Agosti, F. Crestani and G. Pasi eds., Springer Verlag.

[7] D.A. Buell, and D.H. Kraft (1981) Threshold values and Boolean retrieval systems. Information Processing & Management 17, 127-136.

[8] D.A. Buell, (1982). An analysis of some fuzzy subset applications to information retrieval systems. *Fuzzy Sets and Systems*, 7(1), 35-42.

[9] F. Crestani, M. Lalmas, C.J. van Rijsbergen, and I. Campbell (1998). Is this document relevant? …Probably. *ACM Computing Surveys*, 30(4),:528-552, 1998.

[10] F. Crestani and G. Pasi eds. (2000). Soft Computing in Information Retrieval: Techniques and Applications, Physica Verlag, *series* Studies in Fuzziness

[11] F. Crestani and G. Pasi, (1999). Soft Information Retrieval: Applications of Fuzzy Set Theory and Neural Networks. in "Neuro-fuzzy Techniques for Intelligent Information Systems", N.Kasabov and Robert Kozma Editors, Physica-Verlag , Springer-Verlag Group , 287-313, 1999.

[12] P.C. Fishburn, (1970) Utility Theory for Decision Making. New York, John Wiley & Sons.

[13] E. J. Glover, S. Lawrence, M. D. Gordon, W. P. Birmingham, and C. Lee Giles (1999). Web Search -- YourWay. *Communications of the ACM*.

[14] D. Kraft, G. Bordogna, G. Pasi (1999). Fuzzy Set Techniques in Information Retrieval, in "Fuzzy Sets in Approximate Reasoning and Information Systems", J. C. Bezdek, D. Dubois and H. Prade eds, volume della serie "The Handbooks of Fuzzy Sets Series", Kluwer Academic Publishers, 469-510, 1999.

[15] D.H. Kraft, F.E. Petry, B.P. Buckles, and T. Sadasivan, (1997). Genetic Algorithms for Query Optimization in Information Retrieval: Relevance

Feedback, in Sanchez, E., Shibata, T., and Zadeh, L.A. (eds.), Genetic Algorithms and Fuzzy Logic Systems Singapore: World Scientific, 1997.

[16]D. Lucarella and A. Zanzi (1993). Information Retrieval from hypertext: an approach using plausible inference. *Information Processing and Management*, 29(1), 299-312.

[17]S. Miyamoto, S. (1990). *Fuzzy sets in Information Retrieval and Cluster Analysis.* Kluwer Academic Publishers

[18]S. Miyamoto (1990). Information retrieval based on fuzzy associations. *Fuzzy Sets and Systems*, 38(2), 191-205.

[19]A. Molinari, and G. Pasi (1996). A Fuzzy Representation of HTML Documents for Information Retrieval Systems. *In Procedings of the IEEE International Conference on Fuzzy Systems*, 8-12 September, New Orleans, U.S.A., Vol 1, 107-112.

[20]C. D. Paice (1984). Soft evaluation of Boolean search queries in information retrieval systems. *Information Technology: Research Development Applications*, 3(1), 33-41.

[21]G. Pasi, R.A.Marques Pereira (1999). A decision making approach to relevance feedback in information retrieval: a model based on a soft consensus dynamics. *International Journal of Intelligent Systems*, Vol. 14(1), 105-122.

[22]G. Pasi (2002). Modelling the notion of preference in Information Systems, it will appear in the *International Journal of Intelligent Systems*.

[23]J. Rhodes, P. Maes (2000). Just-in-time information retrieval agents. *IBM Systems Journal*, Vol. 39, Nos. 3 & 4, pp. 685-704

[24]G. Salton (1989). Automatic Text Processing - The Transformation, Analysis and Retrieval of Information by Computer, Addison Wesley Publishing Company.

[25]G. Salton and C. Buckley (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5), 513-523.

[26]G. Salton, and M.J. McGill (1983). *Introduction to modern information retrieval.* New York, NY: McGraw-Hill.

[27]T. Sejnowsky (1988). Neural Network learning algorithms. In R. Eckmiller and Ch.v.d.Malsburg, editors, Neural Computers, NATO ASI, 1988.

[28]A. Smeaton (1992). Progress in the application of Natural Language Processing to Information Retrieval tasks. The Computer Journal, 35(3):268-278, 1992.

[29]K. A. Sparck Jones (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11-20.

[30]van Rijsbergen, C. J. (1979). *Information Retrieval. London*, England, Butterworths & Co., Ltd.

[31]P. Vincke (1992), *Multicriteria Decision Aid*, John Wiley & Sons.

[32]L. A. Zadeh (1975). The concept of a linguistic variable and its application to approximate reasoning, parts I, II. *Information Science*, 8, 199-249, 301-357.

[33]L.A. Zadeh (1983). A computational Approach to Fuzzy Quantifiers in Natural Languages, Computing and Mathematics with Applications. 9, 149-184.

[34]R.R. Yager, (1988) On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decision Making, IEEE Trans. on Systems Man and Cybernetics, 18, 1, 183-190.