Fuzzy Sets in Information Retrieval: State of the Art and Research Trends

Gabriella Pasi

Abstract In this contribution some applications of Fuzzy Set Theory to Information Retrieval are described, as well as the more recent outcomes of this research field. Fuzzy Set Theory is applied to Information Retrieval to the main aim to define *flexible* systems, i.e. systems that can represent and manage the vagueness and subjectivity which characterizes the process of information representation and retrieval.

1 Introduction

The advent and rapid diffusion of the Internet and the birth of the World Wide Web have caused a strong resurgence of interest in Information Retrieval, a Computer Science discipline whose roots date to late 60ties. With the diffusion of the World Wide Web the information available on-line have been increasing, and consequently the need for effective systems that allow an easy and flexible access to information has become a urgent need [38]. By flexibility is here meant the capability of the system to both manage imperfect (vague and/or uncertain) information, and to *adapt* its behaviour to the user context. Moreover, more recently, the increasing interest in defining the so called Semantic Web requires the definition of a basic infrastructure more powerful and flexible than the existing one to organise and to give a meaning to the available information, and to allow a better communication between humans and machines.

Search engines represent the most recent outgrowth of IR [20]. However, despite of the above mentioned needs most search engines are based on retrieval models defined several years ago, and, more surprisingly, the query language on which these systems are based is the Boolean query language, defined as the first formal query language for IRSs,. The Boolean query language forces the user to precisely express her/his information needs as a set of un-weighted keywords, thus not allowing users to express vague requirements for specifying selection conditions tolerant to imprecision. Two distinct users formulating the same query will obtain the same results despite of the fact that their choices can be defined on different criteria and to different aims.

For the above mentioned reasons, in recent years a great deal of research has been devoted to the promising direction of improving the (semi) automatic access

H. Bustince et al., (eds.), *Fuzzy Sets and Their Extensions: Representation, Aggregation and Models.*

to information, by modelling the subjectivity, vagueness and imprecision intrinsic in the process of locating relevant information. To this aim the application of Soft Computing techniques has been experienced as a means to obtain a greater flexibility in designing systems for Information Access [24, 18]. 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, and multi-valued logics. As each of these methodologies allows to independently represent imprecision, uncertainty and learning, it is frequently advantageous to employ them in combination, rather than exclusively. Because of these characteristic Soft Computing has provided very powerful tools for modelling flexibility in IRSs.

In particular, Fuzzy Set Theory has been applied to IR starting in the 70ties, and has allowed the definition of retrieval techniques capable of modelling, at some extent, the human subjectivity for estimating the partial relevance of documents to the user needs. The objective of this contribution is to provide an overview of how fuzzy set theory has been applied to the aim of designing flexible Information Retrieval Systems. The chapter is organized as follows: in the next section the Information Retrieval problem is introduced. In Sect. 3 an overview of the main approaches to apply fuzzy set theory to model flexible Information Retrieval Systems is presented. In Sect. 4 a description of the traditional fuzzy document representation is first sketched; then some more recent and promising approaches to fuzzy indexing are described. Section 5 is devoted to the description of flexible query languages for Information Retrieval Systems based on the specification of soft constraints expressed by linguistic selection conditions which capture the vagueness of the user needs and simplify the query formulation.

In Sect. 6 some approaches to the application of fuzzy set theory to distributed information retrieval are described. Finally, in Sect. 7 a description of fuzzy associative retrieval models based either on fuzzy pseudo-thesauri of terms or fuzzy clustering techniques are introduced.

2 Information Retrieval

Information Retrieval (IR) aims at defining systems able to provide a fast and effective content-based access to a large amount of stored information usually organized in documents (information items) [3, 40, 41, 42, 44]. Information can be multimedia: textual, visual, or auditory, although most actual IR systems (IRS) store and enable the retrieval of only textual information.

A user accesses the IRS by explicitly formulating a query through a set of constraints that the relevant information items must satisfy. The aim of the IRS is to evaluate the user query and to retrieve all documents which it estimates relevant to that query. This is achieved by comparing the formal representation of the documents with the formal user's query. The activity of IRSs is then based 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 a user expects from an IRS is a list of the relevant documents ordered according to her/his preferences. The IRS acts then as an intermediary in this decision process: it "simulates" the decision process that the user would personally undertake. The documents constitute the alternatives on which the decision process has to be performed to the aim of identifying the relevant ones [45].

In order to estimate the relevance of each document with respect to a specific user query the IRS must be based on a formal model which provides a formal representation of both documents and user queries. The main components of IRSs are: a collection of documents, a query language which allows the expression of selection criteria synthesizing the users' needs, and a matching mechanism which estimates the relevance of documents to queries (see Fig. 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.

Most of the existing IRSs and search engines offer a very simple modelling of IR, which privileges efficiency at the expenses of effectiveness. A crucial aspect affecting the effectiveness of an IRS 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 keywords specification, and do not allow to express uncertainty and vagueness in specifying the constraints that the relevant information items must satisfy. In real situations, however, the users would find much more natural to express their information needs in an uncertain and vague way.

Another important aspect which affects the effectiveness of IRSs is related to the way in which the documents' information content is 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 that each user looks at a document content in a

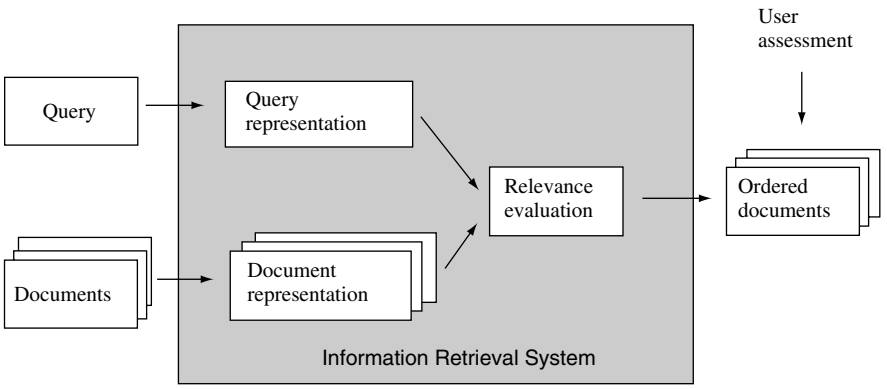

Fig. 1 Scheme of a system for the storage and retrieval of information

subjective way, by possibly 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.

In recent years a strong deal of research to improve IRSs was devoted to the modelling of the concept of partiality intrinsic in the IR process and to making 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 several different directions to the aim of modelling the vagueness and uncertainty that invariably characterize the management of information. In particular, a set of approaches that has received a strong interest goes under the name of Soft Information Retrieval [8, 18, 19, 24]. These approaches apply some of the so called Soft Computing techniques, among which Fuzzy Set Theory.

In particular, 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 synthesized.

3 Applications of Fuzzy Set Theory to Information Retrieval

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

1. to define new IR models;
2. to deal with the imprecision and subjectivity that characterize the document indexing process;
3. to manage the user's vagueness in query formulation;
4. to soften the associative mechanisms, such as thesauri and documents' clustering algorithms, which are often employed to extend the functionalities of the basic IR scheme;
5. to the definition of meta-search engines and to define flexible approaches to distributed IR;
6. to represent and inquiry semi-structured information (XML).

A survey on the definition of fuzzy IR models and of fuzzy generalizations of the Boolean IR model can be found in [8, 27]. Fuzzy generalizations of the Boolean model have been defined to the aim of designing 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 those which are not relevant. 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 [40].

Another approach to fuzzy modelling of IR is based on the use of linguistic information at various levels in the retrieval process [7, 22, 23]. More recently some interesting approaches have been defined to possibilistic based information retrieval [14, 16, 29].

At the level of document indexing some fuzzy techniques have been defined to the aim of providing more specific and personalized representations of documents than those generated by the existing indexing procedures. In Sect. 4, the basic fuzzy interpretation of the weighted document representation is introduced. 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 retrieved document. This value is called the Retrieval Status Value (RSV), and can be used for ranking documents. However, as it will be seen in Sect. 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. In Sect. 4 a fuzzy and personalised 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 [6, 12]. An indexing procedure for HTML documents is also shortly described [31, 35].

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, and by softening the way in which (weighted) search terms can be aggregated. Query languages based on numeric query term weights with different semantics have been first proposed as an aid to define more expressive selection criteria [17, 27]. 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 [4]. Another level of flexibility concerns the definition of soft aggregations of the selection criteria, by means of operators characterized by a parametric behaviour which can be set between the two extremes AND and OR adopted in the Boolean language. In [5, 30] the Boolean query language has been generalized by defining aggregation operators as linguistic quantifiers such *as at least k* or *most of*. In [13] an approach to extend the query languages for inquiring XML documents has been proposed. These extensions are presented in Sect. 5.

Fuzzy associative mechanisms based on thesauri or clustering techniques [33, 34] have been defined in order to cope with the incompleteness characterizing either the representation of documents or the users' queries. Fuzzy thesauri and pseudo-thesauri 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 pseudo-thesauri is to expand 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. These approaches are more extensively explained in Sect. 7.

4 Fuzzy Approaches to Document Indexing

The production of effective retrieval results depends on both subjective factors, such as the users' capability to express their information needs through a formal query, and the characteristics of the Information Retrieval System. A component which plays a crucial role 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: documents are represented by 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 [40, 42]. The vector space model, the probabilistic models and fuzzy models adopt a weighted document representation. 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. In this case an indexing function F computes for each document d and each term t a numeric value. 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 document d, and inversely proportional to the frequency of the term in the documents of the archive:

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

where:

- tf_{dt} is a normalized term frequency which can be defined as: $tf_d = OCC_{dt}/MAXOCC_d$; where OCC_{dt} is 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. 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 [3, 40, 44].

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 normalized weighted document representation and to interpret it as a fuzzy set of terms [8, 27]. 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 be naturally described by adopting the function F (such as the one defined in (1) but normalized so as to obtain values in the range [0,1]) 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.

4.1 A Fuzzy Approach to Personalized Document Indexing

The weighted representation of documents based on function (1) 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 an XML document organized in "logical" sections. For example scientific papers are usually organised into sections like *title*, *authors*, *abstract*, *introduction*, *references*, etc. An occurrence of a term in the *title* has a distinct informative role than an occurrence of the same term in the *references*. Moreover, indexing procedures based on the F function like the one 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 the estimate of relevance of a given document could take advantage from an explicit user's indication of her/his interpretation of the document's structure, and supports the idea of *dynamic* and *adaptive* indexing [6, 12]. 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 could 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 can 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 generate a personalized document representation, which is distinct for distinct users.

In [6] 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 becomes then a fuzzy binary relation defined on the Cartesian product T × 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 Fig. 2). This is done by the specification of a linguistic quantifier such as *at least k* and *most* [47]. In the fuzzy indexing model defined in [6, 12] linguistic quantifiers are formally defined as Ordered Weighted Averaging (OWA) operators [48].

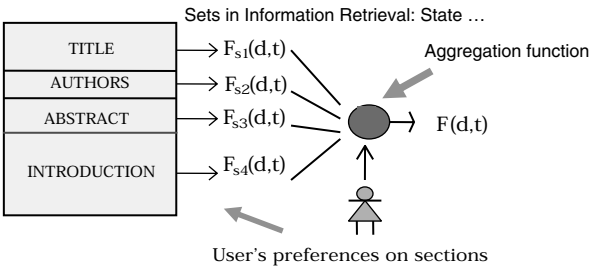

Fig. 2 Sketch of the personalized indexing procedure

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 [31, 35] 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 their 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 occurrences according to the importance of tags in which they appear.

4.2 A Fuzzy Approach to Concept-based Document Indexing and Retrieval

Recently, an increasing number of approaches to IR have defined and designed IR models which are based on concepts rather than keywords, thus modeling document representations at a higher level of granularity, trying to describe the topical content and structure of documents [2]. These efforts gave raise to the so called concept-based Information Retrieval, which aims at retrieving relevant documents on the basis of their meaning rather than their keywords. The main idea at the basis of conceptual IR is that the meaning of a text depends on conceptual relationships to objects in the world rather than to linguistic relations found in text or dictionaries [43]. To this aim, sets of words, phrases, names are related to the concepts they encode.

In [15] a fuzzy set approach to concept-based Information Retrieval has been proposed. Based on the existence of a conceptual hierarchical structure which encodes the contents of the domain to which the considered collection of documents belongs, both documents and queries are represented as weighted trees. The evaluation of a conjunctive query is then interpreted as computing a degree of inclusion between sub-trees. The ontology-based description of the contents of the documents takes into account the semantic equivalences between expressions, as well as the basic principle stating that if a document explicitly heavily includes some terms, it also concerns to some extent more general concepts. This latter point is handled at the technical level by a completion procedure which assesses positive weights also to terms which do not appear directly in the documents. The possible completion of queries is also discussed in [16].

5 Fuzzy Approaches to the Definition of Flexible Query Languages

By flexible query language is intended a language that makes possible a simple and natural 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 consist of either both of the two following soft components or just one: the first component is constituted by weighted terms that are interpreted as flexible constraints on the significance of the index terms in each document representation. The second component is constituted by linguistic aggregation operators which can be applied to the flexible constraints in order to specify 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 linguistic 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 generalized Boolean query language [46]. When flexible 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 flexible constraints, and is defined as a bottom-up evaluation procedure: first, each atomic selection condition (flexible 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. Flexible 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 flexible constraint imposed by the *weight* associated with query term t by the index term weight of t in document d. The result of the evaluation is a fuzzy set: $\Sigma_{d\in D}\mu_{weight}(F(d,t))/d$.

A first proposal to specify flexible constraints was by means of numeric weights associated with terms. A numeric weight identifies a constraint on the weighted document representation, which depends on the considered semantics. Distinct semantics have been associated with query weights [17, 27]. 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 evaluated in a gradual way. To overcome this limitation and to make the query language more user friendly, in [4] a linguistic extension of the Boolean query language was defined, based on the concept of linguistic variable [46]. 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. A pair <t, *important*>, expresses a flexible constraint evaluated by the function $\mu_{important}$ on the term significance values (the F(d,t) values). The evaluation of the relevance of a given document d to a query consisting of the pair <t,*important*> is then computed by applying the function $\mu_{important}$ to the value F(d,t).

A second approach to make the Boolean query language more flexible has concerned the specification of aggregation operators. In the Boolean query language, the AND and OR aggregation operators are used. When the AND is used for aggregating M selection conditions, the satisfaction of all conditions but one is not tolerated, with the consequence that this may cause the rejection of useful items. To face this problem, 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 [5]. 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 behaviour 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, have been defined by Ordered Weighted Averaging (OWA) operators [48]. An alternative approach is proposed in [30].

By adopting linguistic quantifiers, the requirements of a complex Boolean query can be 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 [12] a generalisation of the Boolean query language that allows to personalize the search in structured documents (as showed in Sect. 4) was 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*.

5.1 An Approach to Extend the XPath Query Language

With the development of the World Wide Web the diffusion of the de-facto standards for the definition of structured documents such XML, witnesses the tendency of producing documents in which the information is organized into (often hierarchical) components. In particular, XML is increasingly gaining importance as a standard format for information interchange on the WWW. XML has been employed as a basic model for describing semi-structured data, and it constitutes the basic standard for representing structured documents in IR.

In order to inquiry semi-structured information the need for flexible query languages has soon emerged. In the context of semi-structured databases, by flexible query languages it is substantially meant languages that take into account the lack of a rigid schema of the database, thus allowing to enquiry both data and the type-/schema [1, 12]. In the context of IRSs, modelling flexibility means to take into account the possibility to make explicit a non-uniform structure of the documents when formulating queries.

In [13], fuzzy set theory has been applied to define a flexible extension of the XPath query language to the aim of expressing soft selection conditions on both the documents' structure and contents. XPath is a standard language (www.w3.org/TR/xpath) that allows to write "tree traversal expressions" for selecting XML tree nodes. XPath expressions are also used as selection conditions in the framework of fully-fledged XML query languages. In the last years, much work has been done towards a standard for XML querying and recently the W3C endorsed the XQuery language (www.w3.org/TR/xquery) as a candidate recommendation. Both in XQuery and in XPath a retrieved information item is usually a "node set".

The extensions of XPath proposed in [13] are finalized at:

- fuzzy sub-tree matching to the aim of providing a ranked list of retrieved information items rather than the usual set oriented one;
- use of fuzzy predicates, to the aim of specifying flexible selection conditions;

- fuzzy quantification, to the aim of allowing the specification of linguistic quantifiers as aggregation operators.

The research work presented in this paper constitutes a step towards the more and more increasingly studied problem of inquiring XML documents not only from a structural point of view, but also from a content-based point of view [25].

6 Fuzzy Approaches to Distributed Information Retrieval

With the increasing use of the network technologies, the need of defining distributed applications has emerged. In distributed Information Retrieval, there are two main models: in the first model the information is considered as belonging to a unique, huge database which is distributed but "centrally" indexed for retrieval purposes. This is the model adopted by search engines on the WWW. A second model is based on the distribution of the information on distinct databases, independently indexed, and thus constituting distinct sources of information. This last model gives rise to the so called distributed or multi-source information retrieval problem. In this second case the databases reside on distinct servers each of which can be provided with its own search engine (IRS). The multi-source information retrieval paradigm is more complex than the centralized model as it presents additional problems, such as the selection of an appropriate information source for a given information need. A common problem which can be identified with both models is the problem of list fusion. In the case in which we have a unique, huge and distributed information repository (like in the WWW), and distinct IRSs (search engines), which can be used to inquiry overlapping collections, meta-search engines have been defined to improve the effectiveness of the individual search engines. The main aim of a meta-search engine is to submit the same query to distinct search engines and to fuse the individual resulting lists into an overall ranked list of documents that is presented to the user. In this case we typically have overlapping individual lists since a document may be retrieved by more than a single search engine. The fusion method must then be able to handle situations in which a document may appear in more than one list and in different positions within them. In the case of multi-source information retrieval the problem is to merge the lists resulting from the processing of the same query by (generally distinct) search engines on the distinct databases residing on distinct servers. However, in this second case we generally do not have overlapping lists as a result of the same query evaluation. Typically a document will be retrieved by just one single search engine, and thus the fusion problem is simplified with respect to the previous case. Recently in the literature several papers have addressed the problem of defining effective solutions to the problem of retrieving information on a network. In [10] some approaches to the definition of meta-search engines are presented, while in [9] some solutions to the problem of multi-source information retrieval are described. In this last paper both previous models have been considered, and some fuzzy approaches to the solution of the two above mentioned problems have been proposed. The uniqueness of these

approaches is that they are based on soft computing techniques to more flexibly model the resource selection problem (in distributed information retrieval), and the list fusion problem [9, 10].

In [10] a meta-search model has been proposed where the soft fusion of overlapping ordered lists into an overall ordered list is regarded as a Group Decision Making activity in which the search engines play the role of the experts, the documents are the alternatives that are evaluated based on a set of criteria expressed in a user query, and the decision function is a soft aggregation operator modelling a specific user retrieval attitude.

7 Fuzzy Associative Mechanisms

Associative retrieval mechanisms are defined to enhance the retrieval of traditional IRSs. They work by retrieving additional documents that are not directly indexed by the terms in a given query but are indexed by other terms, associated descriptors. The most common type of associative retrieval mechanism is based on the use of a thesaurus to associate entry terms with related terms. In traditional associative retrieval the associations are crisp.

The fuzzy associative retrieval mechanisms are based on the concept of fuzzy associations [33]. A fuzzy association between two sets $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$ is formally defined as a fuzzy relation $f : X \times Y \rightarrow [0,1]$: the value f(x,y) represents the degree of strength of the association existing between the values $x \in X$ and $y \in Y$.

In information retrieval, different kinds of fuzzy associations can be derived depending on the semantics of the sets X and Y.

Fuzzy associative mechanisms employ fuzzy thesauri, fuzzy pseudothesauri, and fuzzy clustering techniques to serve three alternative, but compatible purposes:

- to expand the set of index terms of documents with new terms
- to expand each of the search terms in the query with associated terms,
- to expand the set of the documents retrieved by a query with associated documents.

A thesaurus is an associative mechanism that can be used to improve both indexing and querying. It is well known that the development of thesauri is very costly, as it requires a large amount of human resources. Moreover, in highly dynamic situations, where terms are added and new meanings derived for old terms quite rapidly, the thesaurus needs frequents updates. For this reason, methods for the automatic construction of thesauri have been proposed, based on statistical criteria such as the terms' co-occurrences, i.e., the simultaneous appearance of pairs (or triplets, or even larger subsets) of terms in the same documents.

In a thesaurus the relations defined between terms are of different type: if the associated descriptor has a more general meaning than the entry term, the relation is

classified as broader term (BT), while a narrower term (NT) is the inverse relation; synonyms or near-synonyms are associated by a related term (RT) relation.

Some authors have proposed the definition of fuzzy thesauri, see [34, 36, 37], where the links between terms are weighted to indicate the strength of the association. Moreover, this notion includes generalizations such as fuzzy pseudo-thesauri [34], and fuzzy associations based on a citation index [36].

7.1 Fuzzy Clustering

Clustering in information retrieval is applied for partitioning a given set of documents D into groups using a measure of similarity (or distance) which is defined on every pairs of documents. The similarity between documents in the same group should be large, while it should be small for documents in different groups. A common method to perform clustering of documents is based on the simultaneous occurrences of citations in pairs of documents. Documents are so clustered using a measure defined on the space of the citations. Generated clusters can then be used as an index for information retrieval; that is, documents which belong to the same clusters as the documents directly indexed by the terms in the query are retrieved. Often, similarity measures are suggested empirically or heuristically [41, 42]. When adopting the fuzzy set model, clustering can be formalized as a kind of fuzzy association. In this case, the fuzzy association is defined on the domain D × D, where D is the set of documents. By assuming R(d) to be the fuzzy set of terms representing a document d with membership function values $\mu_d(t)$=F(d,t) being the index term weights of term t in document d, the symmetric fuzzy relation s is taken to be the similarity measure for clustering documents:

$$\begin{aligned} s(d_1, d_2) &= \sum_{k=1}^{M} \min[\mu_{d1}(t_k), \mu_{d2}(t_k)] / \sum_{k=1}^{M} \max[\mu_{d1}(t_k), \mu_{d2}(t_k)] \\ &= \sum_{k=1}^{M} \min[F(t_k, d_1), F(t_k, d_2)] / \sum_{k=1}^{M} \max[F(t_k, d_1), F(t_k, d_2)] \end{aligned} \quad (2)$$

in which M is the cardinality of the set of index terms T.

In fuzzy clustering, documents can belong to more than one cluster with varying degree of membership [21, 26, 28, 39]. Each document is assigned a membership value to each cluster. In a pure fuzzy clustering, a complete overlap of clusters is allowed. Modified fuzzy clustering, or soft clustering, approaches use threshold mechanisms to limit the number of documents belonging to each cluster. The main advantage of using modified fuzzy clustering is the fact that the degree of fuzziness is controlled.

In [11] a new unsupervised hierarchical fuzzy clustering algorithm has been defined to the aim of identifying the main categories of news in a new-stream

information filtering system. In the following we list the distinguished characteristics of the proposed approach to support category based news filtering.

The output of the proposed fuzzy algorithm is fuzzy hierarchy of the news given as input; this reflects the very nature of a news, which may deal with multiple topics. The algorithm computes a membership degree in [0,1] for each item (news) to each generated fuzzy cluster. This allows to rank the news within a cluster and thus easily support flexible filtering strategies such as the selection of the top ranked news within a cluster of interest. The generated fuzzy hierarchy represents the topics at different levels of granularity, from the most specific ones corresponding to the clusters of the lowest hierarchical level (the deepest level in the tree structure representing the hierarchy), to the most general ones, corresponding with the clusters of the top level. Since topics may overlap one another, the hierarchy is fuzzy, thus allowing each cluster of a level to belong with distinct degrees to each cluster in the next upper level. The proposed algorithm works bottom up in building the levels of the fuzzy hierarchy. Once the centroids of the clusters in a level of the hierarchy are generated, the fuzzy clustering algorithm is re-applied to group the newly identified centroids into new fuzzy clusters of the next upper level. In this way, each level contains fuzzy clusters that reflect topics homogeneous with respect to their specificity (or granularity), so that, in going up the hierarchy, more general topics are identified.

The clusters hierarchy can be easily and efficiently updated on-line when recent news arrive on the stream. This may possibly increase the number of the clusters already identified, and thus may require to compute the association of the old news to the new clusters.

Since the optimal number of clusters to generate is unknown, the proposed algorithm automatically determines this number. The procedure is based on the analysis of the shape of the cumulative histogram curve of overlapping degrees between pairs of news vectors. It identifies the number of clusters of news sharing a minimum overlapping degree corresponding with the point on the curve of highest trend's variation.

8 Conclusions

In this contribution some approaches to the definition of flexible Information Retrieval Systems by applying Fuzzy Set Theory 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). Other promising directions are constituted by conceptual document indexing, and flexible distributed Information Retrieval.

References

1. Abiteboul S., *Querying Semi-Structured Data*, Lecture Notes In Computer Science, Proceedings of the 6th International Conference on Database Theory, pp. 1–18, 199.
2. Azzopardi, L., Girolami M. L., and van Rijsbergen C.J., *Topic Based Language Models for ad hoc Information Retrieval*, in: Proceedings of the International Joint Conference on Neural Networks, Budapest, Hungary, 2004.
3. Baeza-Yates R., Ribeiro-Neto B., *Modern Information Retrieval*. Addison-Wesley, Wokingham, UK, 1999.
4. Bordogna G.and Pasi G., *A fuzzy linguistic approach generalizing Boolean information retrieval: a model and its evaluation*, Journal of the American Society for Information Science, 44(2), pp. 70–82, 1993.
5. Bordogna G.and Pasi G., *Linguistic aggregation operators in fuzzy information retrieva,*. International Journal of Intelligent systems, 10(2), pp. 233–248, 1995.
6. Bordogna G.and Pasi G., *Controlling retrieval trough a user-adaptive representation of documents*, International Journal of Approximate Reasoning, 12, 317–339, 1995.
7. Bordogna G.and Pasi G., *An Ordinal Information Retrieval Model*, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 9, 2001.
8. Bordogna G.and Pasi G., *Modelling Vagueness in Information Retrieval*, in *Lectures in Information Retrieval*, M. Agosti, F. Crestani and G. Pasi eds., Springer Verlag., 2001.
9. Bordogna G., Pasi G., and Yager R.R., *Soft approaches to distributed information retrieval*, International Journal of Intelligent Systems, Vol. 34, pp. 105–120, 2003.
10. Bordogna G.and Pasi G., *Soft fusion of Infomation Accesses*, Fuzzy Sets and Systems, 148, pp. 205–218, 2004.
11. Bordogna G.., Pagani M., and Pasi G., *A dynamical Hierarchical fuzzy clustering algorithm for document filtering*, in "Soft Computing for Information Retrieval on the Web", Springer Verlag, 2006.
12. Bordogna G. and Pasi G., *Personalized Indexing and Retrieval of Heterogeneous Structured Documents*, Information Retrieval, Kluwer, Vol. 8, Issue 2, pp. 301–318, 2005.
13. Braga D., Campi A.. Damiani E.,.Pasi G, Lanzi PL, *FXPath: flexible querying of XML documents*, in Proceedings of EUROFUSE 2002, Varenna, Italy, 2002.
14. Boughanem M., Loiseau Y., Prade H., *Improving document ranking in information retrieval using ordered weighted aggregation and leximin refinement*., in: EUSFLAT-LFA 2005, 4th Conference of the European Society for Fuzzy Logic and Technology and 11me Rencontres Francophones sur la Logique Floue et ses Applications, pp. 1269–1274, 2005.
15. Boughanem M., Pasi G., Prade H., Baziz M., *A fuzzy logic approach to information retrieval using an ontology-based representation of documents*, in "Fuzzy Logic and the Semantic Web" (E. Sanchez, Ed.), Elsevier Science, 2006.
16. Brini A., Boughanem M., Dubois D., *A Model for Information Retrieval Based on Possibilistic Networks*, in: String Processing and Information Retrieval (SPIRE 2005), LNCS, Springer Verlag, pp. 271–282, 2005.
17. Buell D.A., and Kraft D.H., *Threshold values and Boolean retrieval systems*, Information Processing & Management 17, pp. 127–136, 1981.
18. Crestani F. and Pasi G. eds., *Soft Computing in Information Retrieval: Techniques and Applications,* Physica Verlag, *series* Studies in Fuzziness, 2000.
19. Crestani F. and Pasi G., *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 , pp. 287–313, 1999.
20. Glover E. J., Lawrence S., Gordon M. D., Birmingham W. P., and Lee Giles C., *Web Search – YourWay*, Communications of the ACM, 1999.
21. Hathaway R.J., Bezdek J.C., and Hu Y., *Generalized Fuzzy C-Means Clustering Strategies Using Lp Norm Distances*, IEEE Transactions on Fuzzy Systems, 8(5), pp. 576–582, 2000.

22. Herrera-Viedma E., *Modeling the Retrieval Process of an Information Retrieval System Using an Ordinal Fuzzy Linguistic Approach*, Journal of the American Society for Information Science and Technology (JASIST), Vol. 52 N. 6, pp. 60–475, 2001.
23. Herrera-Viedma E., Cordon O., Luque M., Lopez A.G., Muñoz A.N., *A Model of Fuzzy Linguistic IRS Based on Multi-Granular Linguistic Information*, Int. Journal of Approximate Reasoning, 34(3), pp. 221–239, 2003.
24. Herrera-Viedma E., Pasi G. and Crestani F. eds., *Soft Computing in Web Information Retrieval: Models and Applications*, Series of Studies in Fuzziness and Soft Computing, Springer Verlag, 2006.
25. Fuhr N., Lalmas M eds., *Introduction to the Special Issue on INEX*, Information Retrieval, Kluwer, 8(4), pp. 515–519, 2005.
26. Kraft D., Chen J., Martin–Bautista M.J., Vila M.A., *Textual Information Retrieval with User Profiles using Fuzzy Clustering and Inferencing*, in: Intelligent Exploration of the Web, Szczepaniak P., Segovia J., Kacprzyk J., Zadeh L.A. eds., Studies in Fuzziness and Soft Comp. Series, 111, Physica Verlag, 2003.
27. Kraft D., Bordogna G., Pasi G., *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 of the series "The Handbooks of Fuzzy Sets Series", Kluwer Academic Publishers, pp. 469–510, 1999.
28. Lin K., Ravikuma K., *A Similarity-Based Soft Clustering Algorithm for Documents,* in: Proceedings of the 7th International Conference on Database Systems for Advanced Applications, pp. 40–47, 2001.
29. Loiseau Y., Boughanem M., Prade H., *Evaluation of term-based queries using possibilistic ontologies*, in: Soft Computing for Information Retrieval on the Web, Herrera-Viedma E., Pasi G., Crestani F. Eds., Springer-Verlag, 2006.
30. Losada D., Diaz-Hermida F. and Bugarin A., *Semi-fuzzy quantifiers for information retrieval*, in: "Soft Computing in Web Information Retrieval: Models and Applications", Series of Studies in Fuzziness and Soft Computing, Springer Verlag. Edited by E .Herrera-Viedma, G. Pasi and F. Crestani, volume 197/2006.
31. Marques Pereira R.A., Molinari A., and Pasi G., *Contextual weighted representations and indexing models for the retrieval of HTML documents*, Soft Computing, Vol. 9, Issue 7, pp. 481–492, July 2005.
32. Mendes Rodrigues M.E.S. and Sacks L., *A Scalable Hierarchical Fuzzy Clustering Algorithm for Text Mining*, in: Proceedings of the 4th International Conference on Recent Advances in Soft Computing, RASC'2004, pp. 269–274, Nottingham, UK, 2004.
33. Miyamoto S., *Fuzzy sets in Information Retrieval and Cluster Analysis.* Kluwer Academic Publishers, 1990.
34. Miyamoto S., *Information retrieval based on fuzzy associations*, Fuzzy Sets and Systems, 38(2), pp. 191–205, 1990.
35. Molinari A., and Pasi G., *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, pp. 107–112, 1996.
36. Nomoto, K., Wakayama, S., Kirimoto, T., and Kondo, M., *A fuzzy retrieval system based on citation*, Systems and Control, 31(10), pp. 748–755, 1987.
37. Ogawa, Y., Morita, T., and Kobayashi, K., *A fuzzy document retrieval system using the keyword connection matrix and a learning method*, Fuzzy Sets and Systems, 39(2), pp. 163–179, 1991.
38. Pasi G., *Modelling Users' Preferences in Systems for Information Access*, International Journal of Intelligent Systems, Vol. 18, pp. 793–808, 2003.
39. Pedrycz W., *Clustering and Fuzzy Clustering*, Chap. 1, in: Knowledge-based clustering, J. Wiley and Son, 2005.
40. Salton G., *Automatic Text Processing - The Transformation, Analysis and Retrieval of Information by Computer*, Addison Wesley Publishing Company, 1989.
41. Salton G., and McGill M.J., *Introduction to modern information retrieval*. New York, NY: McGraw-Hill, 1983.

42. Sparck Jones K. A., *A statistical interpretation of term specificity and its application in retrieval*, Journal of Documentation, 28(1), pp. 11–20, 1972.
43. Thomopoulos R., Buche P., Haemmerlé O., *Representation of weakly structured imprecise data for fuzzy querying*. Fuzzy Sets and Systems, 140, 111–128, 2003.
44. van Rijsbergen C.J., *Information Retrieval. London,* England, Butterworths & Co., Ltd., 1979.
45. Vincke P., *Multicriteria Decision Aid*, John Wiley & Sons, 1992.
46. Zadeh L. A., *The concept of a linguistic variable and its application to approximate reasoning, parts I, II*, Information Science, 8, pp. 199–249, pp. 301–357, 1975.
47. Zadeh L.A., *A computational Approach to Fuzzy Quantifiers in Natural Languages*, Computing and Mathematics with Applications. 9, 149–184, 1983.
48. Yager, R.R., *On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decision Making*, IEEE Transactions on Systems Man and Cybernetics, 18(1), pp. 183–190, 1988.