Fuzzy Logic Applications for Knowledge Discovery: a Survey

[1]Jorge Ropero, [2]Carlos León, [3]Alejandro Carrasco, [4]Ariel Gómez, [5]Octavio Rivera
[1]*Department of Electronic Technology, University of Seville, jropero@dte.us.es*
[2,3,4,5]*Department of Electronic Technology, University of Seville, ariel@us.es, acarrasco@us.es, cleon@us.es, octavio@dte.us.es*
doi: 10.41256/ijact.vol3.issue6.22

Abstract

In this paper, we present a survey on Fuzzy Logic (FL) applications for Knowledge Discovery (KD), focusing on Information Retrieval (IR) and Information Extraction (IE). KD has been widely used for the search of information in vague, imprecise and noisy environments. Computational Intelligence, and mainly Fuzzy Logic, emerges as an ideal tool for IR and IE systems. We introduce a huge variety of FL applications, using the two main existing approaches: the applications based on the Vector Space Model (VSM); and the applications based on ontologies. For VSM applications, we split the applications into those related to queries, clustering, user profiles and hierarchical relations. Meanwhile, we also consider ontology applications, later focusing in the semantic web. However, all these applications are closely related.

Keywords: *Information Retrieval, Information Extraction, Knowledge Discovery, Computational Intelligence, Fuzzy Logic, Vector Space Model, Ontology*

1. Introduction

The high amount of available information caused by the rise of Information Technology constitutes an enormous advantage for information searchers. Nevertheless, at the same time, a great problem arises as a result of this increase of data: the difficulty to distinguish the necessary information among the whole huge quantity of unnecessary data.

For this reason, Knowledge Discovery (KD), and especially Information Retrieval (IR) and Information Extraction (IE), has hit the scientific headlines strongly in the last times. At first, both arose for document retrieval and extraction, but in the last years its use has been generalized for the search for other types of information, such as the one in a database, a web page or, in general, any set of knowledge. Information Retrieval (IR) is the automatic search of the relevant information contained in a set of knowledge, guaranteeing at the same time that non-relevant retrieved information is as less as possible. Once documents have been retrieved, the challenge is to extract the required information automatically. Information Extraction (IE) is the task in charge of it. IE involves a transformation of a collection of documents, generally helped by an IR system. This collection of documents is transformed into an easier one to assimilate and analyze information. IE tries to extract relevant facts from documents, whereas IR selects relevant documents. However, the edge between IR and IE is not clear, and sometimes they are often used nearly as synonims.

The aim of any system for the access to knowledge is to satisfy user needs on finding the resources that contain the desired knowledge. Nevertheless, the designers of these systems may find some problems [1]:

- User information needs are vague or imprecise.
- User needs may change during his query.
- Users are not conscious of their exact needs of information.
- Information needs are not easy to express in a question in Natural Language (NL).

In this paper, we describe the tools that make the access to the information possible. Given the above mentioned problems, the use of Artificial Intelligence (AI) is desirable. AI is suitable for IR and IE applications for its flexibility. Especifically, we will be limited to one of the branches of AI, Computational Intelligence (CI).

This paper has been organized in four Sections. Section 2 constitutes an introduction to CI and all the techniques that can be applied to Knowledge Discovery, focusing on Fuzzy Logic.

Section 3 introduces the state of the art of FL applications for IE and IR. It is divided in two sections, which correspond to the two main trends nowadays: VSM (Vector Space Model)-based applications and ontology-based applications.

Section 4 shows the main conclusions of this survey.

2. Computational Intelligence

The advance of the electronics and computer science in the last years has been immense, but still there are several tasks that are not solved efficiently. These tasks are usually related with pattern recognition or with systems functioning in noisy and/or vague environments. It is to say, what is usually called *real world*. Man-built machines, nevertheless, obtain better results in tasks related to calculation and logical reasoning. A machine can play magnificent games of chess, but it will be nearly impossible for it to strike a football properly.

In order to solve all these problems, *Computational Intelligence*, (CI) emerged. The most important fields in CI are *Artificial Neural Networks*, (ANN) and *Fuzzy Logic* (FL). Likewise, other techniques, as *Genetic Algorithms* (GA) and *Rough Sets* (RS) theory are also used to lesser extent [2].

Neural networks are based on the introduction of parallel calculation, as it happens in the human brain, and in the processing made by millions of small microcomputers called neurons. Computer networks and neural networks have a great parallelism. Both consist of nodes (computers or neurons), layers and connections. Nevertheless, there are also deep differences between both types of networks. Logically, the power of computation of a computer is many orders greater than the one of a neuron. However, neurons have the advantage of their enormous number and their aptitude to work in parallel with a better performance.

As for Fuzzy Logic, it arises as an answer to the inflexibility of classic binary logic, where the average term does not exist: either it is hot or it is cold [3]. FL gives the possibility of intermediate conditions: cold, cool, warm, hot… Besides, by means of a series of functions, it is possible to provide with a degree of flexibility to these epithets.

FL is then useful for:

- The management of uncertainty.
- The management of the available information, together with uncertainty.

In this approach, thus, precise information and uncertainty are present. Using FL sacrifices a certain amount of precision in favour of uncertainty with the hope of obtaining vaguer, but more robust conclusions. So Fuzzy Logic is an ideal tool for Knowledge Discovery.

Genetic Algorithms (GA) are inspired in biological evolution and its genetic base. These algorithms cause the evolution of a population of individuals, subject to random actions similar to those in the biological evolution (mutations and genetic recombinations). Besides, it is subject to some criteria. Depending on the criteria, the most adapted individuals survive, and the least suitable are rejected. GA are especially used for optimization and can be combined with ANN and FL in order to obtain the ideal parameters of both types of systems.

Finally, Rough Sets (RS) can be used as a theoretical base for some automatic learning problems. They are particularly useful for rule induction and feature selection.

As said in Section 1, it is necessary to consider that the aim of any knowledge access system is to satisfy the needs of the user on accessing to information resources. The existing problems are also mentioned in Section 1.

The habitual query languages are also limited to express the users' needs, being restricted to strict criteria: concepts and logic (AND, OR, NOT). A widely known example of these languages is SQL (*Structured Query Language*), supported by RDBMS (*Relational Database Management Systems*). The problems for non-experts are the following:

- The users must know the terms (attribute names, topic names, etc) used by the knowldege access system.
- The users must be capable of identifying the data that contain relevant information from the data with irrelevant information.
- The users often make mistakes in logical expressions. An example would be to translate "I want information about pop and rock records" for "records AND pop AND rock" when the correct

translation would be "records AND (pop OR rock)". Besides, the graduation of the importance on satisfying different criteria is not possible.

In addition, it is necessary to consider flexibility. For example, in the consultation "I want information about second-hand cars that satisfy a few criteria C in a 10000-12000 euros price range", the price criterion does not necessarily mean that the cars with prices out of this range are completely rejectable, but it means that they are less interesting. For example, a car that costs 8000 or 12500 euros still might turn out to be interesting to a larger or lesser extent, depending on the degree in which they satisfy the criteria C. Another feature to bear in mind is the way of accessing to knowledge on the part of the user. We may distinguish the following ways:

- Query approach: Based in user queries. These queries are the inputs to a system, which must interpret them.
- Inductive approach: based on the information about user interests with regard to the objects in a database. The system must deduce the needs of information behind the user's preferences. This is related to the creation of user profiles

3. Fuzzy Logic Applications for Knowledge Discovery

3.1. Introduction

Search engines, web portals and classic technologies for document retrieval usually consist in searching for keywords in the web. The result may be the finding of thousands of hits, many of them valueless or may be not correct. There are some reasons for this situation:

- The user provides some question syntax, not a meaning. The meaning must be guessed by the system.
- The user is anonymous. The system does not know anything about his mental structure and is incapable of deducting any unknown previous feature.
- Homepages are usually full of advertisements or banners, instead of summaries.
- There is quite a lot of secondary information.

The aim of an IR system is evaluating the degree of importance of the available documents regarding the user queries and recovering the documents with a high degree of satisfaction for the user. To obtain a good functioning, the response to the user must be capable of answering properly to what the user requests. Many of the IR systems were traditionally based on the Boolean logical model. This model considers that user queries may be characterized accurately by a few index terms. Nevertheless, this supposition is inadequate due to the fact that user queries may not be clear enough. The reason for this is that the user may not know very much about the object to search or may not be acquainted with the IR system. To be able to handle the imprecision and the uncertainty in the representation of concepts and words in the real world, most of the models of automatic learning have been related to FL. These FL models overcome the problems created by the abrupt separations between the values of the attributes, providing a soft transition and a good precision compared with constant attributes.

There are several approaches at the moment for information handling in an IR system. One of them is based on the Vector Space Model (VSM) and the other one is related to the concepts of ontology and semantic web. In Figure 1, it is possible to see a conceptual scheme on FL applications for IR.

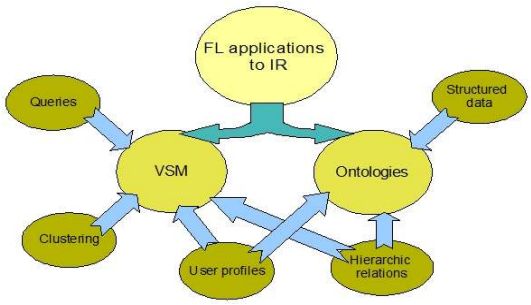

Figure 1. Conceptual scheme on FL applications for IR.

3.2. VSM-based applications

Within VSM-based applications for IR, concepts as queries, clustering, user profiles, or hierarchical relationships take importance.

Queries are the most important application of the Vector Space Model. [4-7]

An example of the use of queries may be found in Haase et al's work [4]. An internet portal is developed by some Austrian investigation centers and leads to databases, electronic and printed documents, as well as to queries and services. The database is based on objects of knowledge: the content of a document or the service is defined by a two-dimensional matrix (map of knowledge) based on a scheme of decimal classification and a game of attributes, and it is accessed by means of queries. These queries are interpreted as linguistic variables that can be used to construct a "map of questions". In the same way, Cordon et al. [5] introduce the so-called persistent queries, which are specific questions used in IR systems to represent long-term user needs. These queries may have many different structures. The so-called bag-of-words model is the most widely used. The bag-of-words model is a simplified supposition used in Natural Language Processing (NLP) document retrieval. In this model, a text is represented as an untid collection of words, disregarding the grammar and even the order of the words. In Cordon et al.'s work, there is an attempt of achieving a more representative structure for the search of texts. With this aim, soft computing techniques are used: FL is used for inference and representation, improving the existing Boolean techniques.

Mercier & Beigdeber are in a similar line [6]. They are based on the idea that the nearer the terms in a document are, the more relevants these terms are. With this idea, a new IR method is tested. This method is based on the fuzzy degree of proximity of the occurrences of the desired term to calculate its importance in a document. The fuzzy proximity of a term is controlled by a "function of influence". Given a term and a document, the function of influence associates every position in the text with a value dependent on the distance to the nearer occurrence of the same term.

Though the major application of fuzzy logic to IR is text and document retrieval, there are also other applications for queries, like Content Based Image Retrieval (CBIR) [7, 8]. Given an image, the system first includes it in a group of regions based on a space of characteristics considers features of colour and texture. A modified k-means algorithm is used. It consists in grouping the objects in a collection in k groups. In every segmented region, fuzzy logic is applied in order to define the features of colour, texture and shape, respectively. The motivation to incorporate fuzzy logic is to locate the imprecision of a typical representation of characteristics to improve robustness and efficiency of the index scheme.

Lastly, our group of investigation has worked with FL in IR for intelligent agents [9]. We developed an intelligent agent that is capable of answering to the needs of the users in their process of retrieving the desired information when it is enormous, heterogeneous, vague, imprecise, or out of order. We created a general method for retrieving and extracting information based on the use of FL, as it is an ideal tool for the management of this kind of vague and heterogeneous information. Besides, this method has been implemented and validated for IE in web portals, provided that web portals are a clear example of imprecise and disordered information. This method has been applied in the development of the University of Seville web Intelligent Agent, which is to be functioning soon.

Other Intelligent Agents have also used the VSM [10]. Some VSM applications for queries, including Face Recognition [11] or Term Weighting [12] are categorized in Table 1.

Table 1. Vector Space Model applications for queries

VSM application	***Authors***
Document retrieval based on persistant queries	Cordon, de Moya and Zarco (2004)
Access to databases from an internet portal	Haase, Steinmann and Vajda (2004)
Bird IR using fuzzy semantic concepts	Huang and Tsai (2005)
Document retrieval based on fuzzy proximity	Mercier and Beigbeder (2005)
Content Based Image Retrieval	Yap and Wu (2005)
Intelligent Agents	Kim, Hong and Cho (2007)
Intelligent Agents for web portals	Ropero, Gomez, Leon and Carrasco (2007)
Term Weighting	Ropero, Gomez, Leon and Carrasco (2009)

On the other hand, clustering is used in other works [13-17]. Many of the existing methods for document clustering are based on the VSM classical model, which represents every document as a vector of key terms or key phrases of a fixed size. In big and diverse collections documents, as the World Wide Web, this approximation suffers from an enormous computational overload, since the constant size of the term vectors is equal to the total number of index terms in all the documents.

In Friedman et al. [13], a fuzzy approach for document clustering is proposed. Documents are represented by vectors of variable size. Provided that a document generally contains only a small subset of the terms in the system, the term matrix associated with a document is usually very little dense, with nearly 99% of the components equal to zero.

A set S of *n* vectors with variable length is considered. A weight w_i is assigned to every term t_i. This weight is calculated by a frequency-based indexing model or by a key phrase extraction algorithm. The objetive is to divide S into several clusters, as many as necessary. The final number of clusters is determined by the algorithm and the structure of S. The only demand is that every cluster must include similar vectors, whereas those which belong to different clusters will be radically different. Algorithms of clustering have been developed, based on the local weight and the measure of the similarity of the cosine. When using the cosine method for clustering, it is necessary to define a central cluster *c*, which is the normalized sum of all the vectors in cluster C. Then the inner product of the input vector and the central cluster vector is calculated. It only considers the index terms that appear in both vectors.

Another interesting application from the point of view of the investigation in this field is Subasic and Huettner's work. They combine NLP and FL to analyze the affective content of a text [14]. A main aspect in their system is what they call "set of affection", a set of semantic fuzzy categories based on positive and negative features which is useful to group the documents or texts in diverse clusters based on the affective characteristics of them.

Moreover, fuzzy semantic concepts have been used widely [15, 16]. In Horng et al. [15], an agglomerative hierarchical clustering method is introduced. This method overcomes the main disadvantage of the classical method, where a document cannot belong to several clusters at the same time. There are a few clustering hierarchical algorithms: Hierarchical clustering may build (agglomerative clustering), or divide (divisive clustering) a group of clusters. The classical representation of this hierarchy is a tree (called dendogram) with individual elements in every ending and an only cluster that contains every element in the other one. Agglomerative algorithms start in the leaves of the tree, whereas divisive algorithms start in the root (Figure 2).

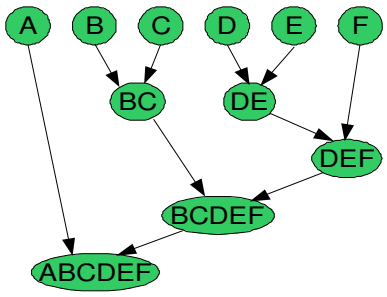

Figure 2. Agglomerative hierarchical clustering algorithm.

Both the proposed method of fuzzy hierarchical clustering as the k-means method have the advantage of their flexibility, for they allow a document to belong to multiple clusters at the same time. The difference between both methods is that using the k-means method forces the number of clusters to be predefined, whereas the fuzzy proposed method uses a "threshold of similarity" α and a "threshold of difference" λ to carry out an automatic process of clustering. Since the hierarchical classical agglomerative method is an abrupt method of clustering (and not a fuzzy one), and every document only can belong to exactly a group, flexibility does not exist.

Rios et al. also introduce a conceptual approach to improve the content of a web site. WUM (Web Usage Mining) tecniques study the user's interests when they browse through the web in order to refine IR according to the user´s preferences [18]. Nevertheless, many investigations in the IR area forget the semantic information also present in the web pages. Therefore, they propose the combination of the so-

called Concept-Based Knowledge Discovery in Text and the study of the visitors' sessions to performance a task of customization. This way, it is possible to obtain the aim of the users who are browsing through a web site. Moreover, it is possible to advice users for a better browsing and to help web site managers to improve web content. In addition, this idea is tested on a real web site to show its efficiency. The solution is based on the idea of finding the relationship between the concepts and the documents. It may be represented as a fuzzy composition of concepts, index terms and words.

A list of concepts and terms that represent these concepts is defined. Nevertheless, it still would be necessary to establish the values of the weights for this relationship. A direct method is used by an expert to define them and, in addition, a simple model which uses the relative frequency of words in every document. This model defines the second fuzzy relation (Terms × Words). The documents were preprocessed to eliminate the HTML and JavaScript codes and a list of stop words was used for eliminating the words that are not important, trying to support nouns, adjectives and verbs. However, stemming was not used. Lastly, the matrix (Concepts × Words) must be obtained. Every row is a concept, every column is a web page and every value of the matrix represents the possibility that a concept is present in a web page.

On the other hand, it is necessary to transform web pages to a more useful way, so the VSM is used to represent every document as a vector of words using the TF-IDF method to establish the weight of every word for every document. It is also necessary to preprocess the inputs (logs) to the web servers in order to understand user's sessions. After this step, the visited pages and the time the visitors browse through them are available. Afterwards, a process of generalization is applied. A Self-Organised Map (SOFM) is used, since it is a non-supervised algorithm. It is not neccessary to know the number of clusters previously. Eventually, groups ordered by their conceptual meaning are obtained. Therefore, a conceptual classification of the documents is carried out on every user session. In last instance, this fact helps the webmasters to improve the usefulness of their web sites. Self-Organised maps were also used by Garcia-Plaza et al. for web page clustering [19].

Table 2 shows a summary of some VSM applications for clustering.

Table 2. Vector Space Model applications for clustering

VSM application	***Authors***
Affective content of a text by semantic fuzzy categories	Subasic and Huettner (2001)
Textual IR	Kraft, Chen, Martin-Bautista and Vila (2003)
Document retrieval in big data collections with lower computational load	Friedman, Last, Zaafrany, Schneider and Kandel (2004)
Document clustering and retrieval – agglomerative clustering -	Horng, Chen, Chang and Lee (2005)
Extraction of concept-based knowledge in a web site	Rios, Velasquez, Yasuda and Aoki (2006)
Document clustering	Cao, Do, Hong and Quan (2008)
Web page clustering	Garcia-Plaza, Fresno and Martinez (2008)
Concept-based knowledge discovery & study of user sessions	Tao, Hong and Su (2008)

The user's profile use is so a paradigm of common use in recent investigations [20-24]. For example, Moradi et al. propose the customization of an IR system results with the aim of bringing a better service to the users according to their profiles, focusing on the particular interests of individual users [21]. With this method, both pages and user's profiles appear as extended fuzzy concept networks. A concept network includes nodes and directed links, where every node represents a concept or a document; every directed link joins two concepts or goes from a concept Ci to a document d_i, labelled with a value between 0 and 1. (Figure 3).

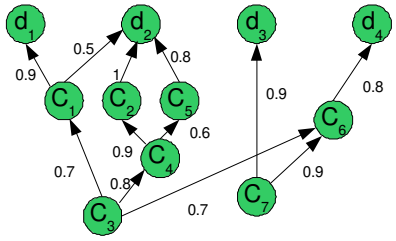

Figure 3. Fuzzy concept network.

Extended fuzzy concept networks are more general. There are tour types of fuzzy relationships between concepts:

- Positive fuzzy association: relates the concepts that have a Fuzzy similar meaning in some contexts. (For example, person → individual).
- Negative fuzzy association: relates the concepts that are fuzzy complementary antonyms. (For example, male → female, big → small).
- Fuzzy generalization: fuzzy generalization of a concept if it incluyes another concept. (For example, vehicle → car).
- Fuzzy specialization: it is the inverse of Fuzzy generalization, it is to say, a concept is a part of another one. (For example, car → vehicle).

In an extended fuzzy concept network, C_1, C_2..., C_7 are concepts and d_1, d_2, d_3 y d_4 are documents. For example, a document d_i may have 50% of concept C_i, 80% of concept C_j and 100% complementary with concept c_k.

Let's suppose that there are two users, A and B. The interests of user A include computer science whereas the interests of user B are more focused on biology. Let's also suppose that they are searching for the same word 'java' on the IR system: user A would prefer documents that are more related to the computer science, whereas user B would prefer documents more related to biology. Basically, the system should modify the weight in the correspondent matrix in order to fit results with every user's profile.

User profile applications of the VSM are shown in Table 3. It is noticeable that Kraft et al. combine user profiles and clustering for textual IR. Martin-Bautista et al. have continued their work on building user profiles with fuzzy logic, as may be seen in papers like [25].

Table 3. Vector Space Model applications with user profiles

VSM application	*Authors*
Textual IR	Kraft, Chen, Martin-Bautista and Vila (2003)
Recommendation system for e-commerce	Cao and Li (2007)
Building user profiles	Martín-Bautista, Vila-Miranda and Escobar-Jeria (2009)
Use of fuzzy concept networks for IR	Moradi, Ebrahim and Ebadzadeh (2008)
Concept-based knowledge discovery & study of user sessions	Tao, Hong and Su (2008)
Weblog extraction	Portmann (2009)

A hierarchical structure also helps Vector Space Model to obtain better results in IR [9, 16, 17]. In fact, the frontier between using a hierarchical structure and ontologies constitutes a thin line [26]. For example, intelligent agents improve their retrieval by using it [9, 26]. Table 4 shows some applications that use a hierarchical structure for improving the results obtined by the VSM.

Table 4. Vector Space Model applications using a hierarchical structure

VSM application	*Authors*
Bird IR using fuzzy semantic concepts	Huang and Tsai (2005)
Document clustering	Cao, Do, Hong and Quan (2008)
Improving the semantic capabilities of Web search engines	Olivas, Garces, de la Mata, Romero and Serrano-Guerrero (2008)

3.3. Ontology-based applications

The other posible approach using FL for IR is ontology-based processing. This model processing may help users to access the information stored in structured or semi-structured documents rather efficiently.

An ontology represents the knowledge in a domain in a structured way and, increasingly, it is being accepted as a key technology when the most important concepts and their mutual relations are stored to provide a shared and common understanding of a domain across certain uses. An ontology is a conceptualization of a domain in a human understandable format, but legible for a machine. It consists of classes, attributes, relations, axioms and entities [27]. Since an ontology describes a domain of interest in an unequivocal way, the schemes of IE based on ontology may help in the elimination of ambiguity in texts in natural language. In addition, it makes an effective semantic analysis of texts. Nevertheless, ontology-based text processing has not been still exploited enough.

In Abulaish & Dey [28], a tecnique using text mining along with a set of ontology concepts is proposed. Text mining extracts fuzzy relationships. Membership values for the relationships are functions of co-occurance of concepts and relationships. The work was done upon the GENIA corpus and it was shown that fuzzy relationships can be used for guided Medline document information extraction.

An intelligent IE system is proposed. There are two main goals. Firstly, the system uses a biological existing ontology (GENIA in this case) for IE in text documents. Secondly, the extracted information is used to evolve from the inflexible existing ontology structures to fuzzy ontology structures, which can assume interconceptual relationships with different degrees. The ontology-based text mining system uses pattern recognition in order to find fuzzy relationships that join ontology concepts from text documents.

The work of Quan et al. is also ontology-based. [29]. In this case, it deals with the development of a semantic help-desk system to support the customer service centre in a semantic web environment. Particularly, a, Formal Concept Analysis (FCA) is developed for fuzzy ontology automated generation. These ontologies may deal with uncertain information. The proposed ontology automated fuzzy generation technique follows these steps: FCA; fuzzy conceptual clustering and ontology generation. In the area of the semantic webs, an ontology is adopted as a standard for knowledge representation. The programs can use the knowledge of the semantic webs to semantically process the information. This way it is allowed that different machines or models produced by different manufacturers could be shared and integrated. In addition, the generated ontology may be used to provide an interpretation on the common faults.

Data can be stored in an unstructured, semi-structured or totally structured way (for example, textual documents or databases). However, implementing an ontology manually is a hard and difficult task. Recent research try to approach this problem by means of the study of free text ontology, semiestructured data, (for example, HTML or XML), or structured data in a database. Different technologies have been used for ontology generation, more remarkably the ones that are based on clustering. Nevertheless, the conceptual formalism supported by a standard ontology may not be sufficient to represent the uncertain information that can be usually found in the information stored in most of databases. For example, in a customer service centre database, every registered problem is described by means of a text stream. A few words can be extracted from it and used for IR. Nevertheless, it is inadequate to treat all the keywords in the same way, provided that some keywords may be more significant than others. A possible way of handling this aspect is the incorporation of fuzzy logic to the ontology. A concept can be described by a series of keywords, whereas fuzzy formal concepts are organized then as a fuzzy concept structure. This structure provides hierarchical relations between the fuzzy formal concepts. Bearing in mind such hierarchical relations, the fuzzy formal concepts can be represented as subconcepts and superconcepts of others. A formal concept is a conceptual model that potentially represents a real concept in the domain of a fuzzy concept. The hierarchical relationships between formal concepts imply a relationship between the generalization and specification of these concepts. A superconcept is more general than its correspondent subconcept and viceversa.

Zhai et al. [30] use FL and ontologies for e-commerce. Since the process for IR is based on the knowledge ontology, the semantic and concept research can be achieved. For instance, a product may be found through some features like the price ("cheap" or "very expensive", for example), the type of customer ("golden customer") or the customer's age ("young", "middle-aged").

A last example of the application of FL to IR is in Zhai et al. [31]. IR is an important aspect Supply Chain Management (SCM). To achieve a semantic fuzzy retrieval, an environment of fuzzy ontologies is applied to the IR in SCM. Hereby, the aim is to retrieve information from, for example, a product, from features like the price of the product, the income of the clients, their conditions or the area of influence of the company.

Table 5 shows a huge variety of ontology-based applications using FL, including such fields as improving the semantic capabilities of Web search engines [26], crisis prediction [32], managing imprecise sets of knowledge [33] and real-time control [34].

Table 5. Ontology-based applications using FL

Ontology application	***Authors***
Document retrieval using normalized collections	Abulaish and Dey (2005)
Semantic help-desk for customer desk	Quan, Hui and Fong (2006)
Crisis prediction	Delavallade, Mouillet, Bouchon-Meunier and Collain (2007)
Improving the semantic capabilities of Web search engines	Olivas, Garces, de la Mata, Romero and Serrano-Guerrero (2008)
Supply Chain Management	Zhai, Wang and Ly (2008)
Fuzzy ontology for e-commerce	Zhai, Liang, Yu and Jiang (2008)
Fuzzy control	Tran (2009)
Fuzzy knowledge management	Lai, Huang, Wu and Chen (2010)

However, the main application of ontologies until now is the so-called semantic web. The semantic web is an extension of the current web provided with meaning, that is to say, a space where the information would have a well defined meaning, so that it could be interpreted both by human agents and by automatic agents. The key to reach this semantic web is the use of metadata to annotate documents; thus, software agents would be able to search, locate, discover, or link documents better than today lexical-based search engines. The development of the semantic web needs the utilization of other languages, as the structured languages XML (Extensible Markup Language) and RDF (Resource Description Framework). These languages may provide of logic and meaning to every page, every file and every resource or content of the network. It also allows computers knowing the meaning of the information. Morever, they must be standardized and formalized for their universal use all around the web. Thus, the use of OWL (Ontology Web Language) constitutes one more step in the consecution of the Semantic Web. In order to create a semantic web, it is necessary to build an ontology or library of descriptive/semantic vocabularies. These vocabularies are defined in RDF format and located in the web to determine the contextual meaning of a word, by means of the consultation to the appropriate ontology. This way, intelligent agents and autonomous programs might search the web automatically and retrieve the pages referred to a term with its precise meaning and concept exclusively. In addition, there are a few semantic web search engines, which search on ontologies and semantic languages like RDF and OWL. Most known is Swoogle, developed by the University of Maryland (Baltimore, USA). It has an index of more than 10.000 documents and terms written in RDF as OWL format, that is to say, it searches for Semantic Web Documents (SWDs).

As FL is an ideal tool to manage imprecise and vague information, it has been widely used in the semantic web field [35]. For example, Laurent et al. [36], build what they call mediator schemas with large amounts of heterogeneous data in XML format. They work with a hierarchical structure using FL to determine the degree of inclusion of a XML document in a determined tree. A tree is no more included or not in another one, but gradually included within it.

Other authors have used OWL for its use in the semantic web. Stoilos et al. [37] provide a fuzzy extension to the OWL web ontology language, using the semantics and abstract syntax of fuzzy OWL to manage uncertainty. Otherwise, Zhai et al. use RDF to represent ontologies, applying them to the fuzzy control of an industrial washing machine [38].

Other uses combining FL and semantic web include intelligent tutoring systems for education [39], IR for biomedical sciences [40], IR in digital libraries [41], product data management [42], web or content distribution [43].

Table 6 shows semantic web applications using FL.

Table 6. Semantic web applications using FL

Ontology application	*Authors*
Fuzzy semantic web using OWL	Stoilos, Stamou, Tzouvaras, Pan and Horrocks (2005)
Fuzzy semantic web using XML	Laurent, Teisseire and Poncelet (2006)
Web content distribution	Borzemski, K. Zatwarnicki and A. Zatwarnicka (2007)
Fuzzy control using RDF	Zhai, Liu, Liang and Jiang (2008)
IR in digital libraries	Martin and Leon (2009)
IR for biomedical domain	Morales-del-Castillo, Peis and Herrera-Viedma (2009)
Integration of semantic web sevices and multi-agent systems	Hemayati, Mohsenzadeh, Seyyedi and Yousefipour (2010)
Intelligent tutoring for education	Milosevic, Sukic and Sendelj (2010)
Product data management	Zhai, Li and Chen (2010)

4. Conclusions

Knowledge Discovery, especially Information Retrieval and Information Extraction, has been widely used for the search of information in vague, imprecise and noisy environments. Computational Intelligences emerges as a solution to all the existant problems and, mainly Fuzzy Logic. Fuzzy Logic is a very useful tool for managing the uncertainty of the avalaible information.

There are two main streams for handling information in Information Retrieval and Information Extraction systems: based on the Vector Space Model, where concepts like clustering, quering, user profiles or hierarchical relationships take importance; and based on ontologies, where the structure of information is essential.

One of the aims of this paper is to present some of the huge variety of applications that may be found in the state of the art. The other objective of the paper is to discuss whether it is convenient to use Vector Space Model or an ontology-based model for Knowledge Discovery.

In the authors' opinión, VSM offers great advantages when the information is not structured or it is structured in a heterogenous way. Tools like Fuzzy Logic may be interesting for managing possible ambiguities. Moreover, creating user profiles and using hierarchic structures help reducing these ambiguities. Otherwise, ontologies represent the knowledge in a domain in a structured way and it is being accepted as a key technique when concepts are important and their mutual relations are stored to provide a shared and common knowledge for certain uses. In domains of knowledge where information is structured this way, the use of ontologies is highly recommendable.

Anyway, all these applications have in common with our proposed work two basic aspects: the great amount of information to handle; and a hierarchical structure or the possibility of clustering the information. Therefore, suitability of Fuzzy Logic for Knowledge Discovery is beyond all doubt.

5. References

[1] Henrik Legind Larsen. "An Approach to Flexible Information Access Systems using Soft Computing". In Proceedings of the 32nd Annual Hawaii International Conference on System Sciences, HICCS99, pp. 231, 1999.

[2] Cheruiyot Wilson, Guan-Zheng Tan, Joseph Cosmas Mushi, Felix Musau, "Genetic Algorithm-Enhanced Retrieval Process for Multimedia Data", IJACT, Vol. 3, No. 3, pp. 153 ~ 167, 2011.

[3] Lofti A. Zadeh, "Fuzzy logic, neural networks and soft computing", Communications of the ACM, vol. 3, no. 3, pp. 77-84. 1994.

[4] Volkmar H. Haase, Christian Steinmann, Stephan Vejda, "Access to Knowledge: Better Use of the Internet". In Proceedings of the Informing Science + IT Education Conference, IS2002, pp. 618-627, 2002.
[5] Oscar Cordon, Felix de Moya, Carmen Zarco, "Fuzzy logic and multiobjective evolutionary algorithms as soft computing tools for persistent query learning in text retrieval environments", In Proceedings of the IEEE International Conference on Fuzzy Systems, vol.1, pp. 571-576, 2004.
[6] Annabelle Mercier, Michel Beigbeder, "Fuzzy Proximity Ranking with Boolean Queries". In Proceedings of the 14th Text REtrieval Conference (TREC 2005), pp. 433-442, 2005.
[7] Kim-Hui Yap, Kui Wu, "A soft relevance framework in content-based image retrieval systems", IEEE Transactions on Circuits and Systems for Video Technology, vol. 15, no. 12, pp.1557-1568. 2005.
[8] Aasia Khanum, "An Intelligent Framework for Natural Object Identification in Images", IJACT, Vol. 2, No. 2, pp. 122 ~ 129, 2010.
[9] Jorge Ropero, Ariel Gomez, Carlos Leon, Alejandro Carrasco, "A method for the access to the contents in a set of knowledge using a fuzzy logic based intelligent agent". In Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2007, vol. 4, pp. 103-108, 2007.
[10] Kyoung Mahn Kim, Jin Hyuk Hong, Sung Bong Cho, "A semantic Bayesian network approach to retrieving information with intelligent conversational agents". Information Processing Management, vol. 43, no. 1, pp.225-236, 2007.
[11] Virendra Vishwakarma, Sujata Pandey, M.N. Gupta, "Fuzzy based Pixel wise Information Extraction for Face Recognition". International Journal of Engineering and Technology, vol. 2, no. 1, pp. 117-123, 2010.
[12] Jorge Ropero, Ariel Gomez, Carlos Leon, Alejandro Carrasco, "Term Weighting: Novel Fuzzy Logic Based Method vs. Classical TF-IDF Method for Web Information Extraction", In Proceedings of the 11th International Conference on Enterprise Information System, pp. 130-137. 2009.
[13] Manahem Friedman, Mark Last, Omer Zaafrany, Moty Schneider, Abraham Kandel, "A new approach for fuzzy clustering of Web documents", In Proceedings of the IEEE International Conference on Fuzzy Systems, vol.1, pp. 377-381, 2004.
[14] Pero Subasic, Alison Huettner, "Affect Analysis of Text Using Fuzzy Semantic Typing". IEEE Transactions on Fuzzy Systems, vol. 9, no. 4, pp. 483-496, 2001.
[15] Yih-Jen. Horng, Shyi-Ming Chen, Yu-Chuan Chang, Chia-Hoang Lee, "A new method for fuzzy information retrieval based on fuzzy hierarchical clustering and fuzzy inference techniques". IEEE Transactions on Fuzzy Systems, vol. 13, no. 2, pp. 216-228, 2005.
[16] Yo-Ping Huang, Tienwei Tsai, "Bird information retrieval using fuzzy semantic concepts". IEEE Potentials, vol. 24, no.3, pp. 26- 28, 2005.
[17] Tru H. Cao, Hai T. Do, Dung T. Hong and Tho T. Quan. "Fuzzy named entity-based document clustering". IEEE International Conference on Fuzzy Systems, pp.2028-2034, 2008.
[18] Sebastian A. Rios, Juan Diego Velasquez, Hiroshi Yasuda, Terumasa Aoki, "Improving the web site text content by extracting concept-based knowledge". Lecture Notes in Artificial Intelligence, vol. 4252, no. 1, pp. 371-378, 2006.
[19] Alberto P. Garcia-Plaza, Victor Fresno, Raquel Martinez, "Web Page Clustering Using a Fuzzy Logic Based Representation and Self-Organizing Maps", In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 851-854, 2008.
[20] Yu-Hui Tao, Tzung-Pei Hong, Yu-Ming Su, "Web usage mining with intentional browsing data", Expert Systems Applications, vol. 34, no. 3, pp. 1893-1904, 2008.
[21] Parham Moradi, Mohammadreza Ebrahim Shiri, Mohammad Mehdi Ebadzadeh, "Personalizing Results of Information Retrieval Systems Using Extended Fuzzy Concept Networks", 3rd International Conference on Information and Communication Technologies: From Theory to Applications, pp.1-7, 2008.
[22] Edy Portmann. "Weblog Extraction with Fuzzy Classification Methods" Second International Conference on the Applications of Digital Information and Web Technologies, pp. 411-416, 2009.
[23] Donald H. Kraft, Jinhua Chen, Maria J. Martin-Bautista and Maria-Amparo Vila, "Textual information retrieval with user profiles using fuzzy clustering and inferencing", Studies In Fuzziness And Soft Computing, pp. 152-165, 2003.

[24] Yukun Cao and Yunfeng Li, "An intelligent fuzzy-based recommendation system for consumer electronic products", Expert Systems with Applications, vol. 33, no. 1, pp. 230-240, 2007.
[25] Maria J. Martín-Bautista, Maria-Amparo Vila-Miranda, Victor H. Escobar-Jeria, "Obtaining User Profiles Via Web Usage Mining", IADIS European Conf. on Data Mining, pp. 73-76, 2008.
[26] Jose A. Olivas, Pablo J. Garces, Javier de la Mata, Francisco P. Romero, Jesus Serrano-Guerrero, "Conceptual soft-computing based web search: FIS-CRM, FISS metasearcher and GUMSe architecture", Studies in Fuzziness and Soft Computing, vol. 218, no. 1, pp. 107-134, 2008.
[27] Nicola Guarino, Pierdale Giaretta. "Ontologies and Knowledge Bases: Towards a Terminological Clarification", Towards Very Large Knowledge Bases: Knowledge Building and Knowledge sharing, N. Mars (ed.) IOS Press, Amsterdam, pp. 25-32, 1995.
[28] Muhammad Abulaish, Lipika Dey, "Biological Ontology Enhancement with Fuzzy Relations: A Text-Mining Framework", Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 379-385, 2005.
[29] Thanh Tho Quan, Siu Cheung Hui, Alvis Cheuk M. Fong, "Automatic fuzzy ontology generation for semantic help-desk support", IEEE Transactions on Industrial Informatics, vol. 2, no. 3, pp. 155-164, 2006.
[30] Jun Zhai, Yiduo Liang, Yi Yu and Jiatao Jiang. "Semantic Information Retrieval Based on Fuzzy Ontology for Electronic Commerce", Journal of Software, vol. 3, no. 9, pp. 20-27. 2008.
[31] Jun Zhai, Qinglian Wang, Miao Lv, "Application of Fuzzy Ontology Framework to Information Retrieval for SCM". Proceedings of the International Symposiums on Information Processing, pp.173-177, 2008.
[32] Thomas Delavallade, Laure Mouillet, Bernardette Bouchon-Meunier and Emmanuel Collain, "Monitoring Event Flows and Modelling Scenarios for Crisis Prediction: Application to Ethnic Conflicts Forecasting", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 15 (Supplement-1), pp. 83-110, 2007.
[33] Lien Fu Lai, Liang-Tsung Huang, Chao-Chin Wu, Shi-Shan Chen, "Fuzzy Knowledge Management through Knowledge Engineering and Fuzzy Logic", Journal of Convergence Information Technology, vol. 5, no. 3, pp. 7-15, 2010.
[34] Chi Lang Tran. "Fuzzy control based on "true and false" philosophy for mechatronics systems", Computer Assisted Mechanics and Engineering Sciences, vol. 16, no. 1, pp. 11-20, 2009.
[35] Mohammad Sadeghzadeh Hemayati, Mehran Mohsenzadeh, Mir Ali Seyyedi and Amin Yousefipour. "A framework for integrating web services and multi-agent systems". 2nd International Conference on Software Technology and Engineering, vol.2, pp 314-319, 2010.
[36] Anne Laurent, Monique Teisseire, Pascal Poncelet, "Fuzzy Data Mining for the Semantic Web: Building XML Mediator Schemas", Volume 1 of Capturing Intelligence, pp. 249-265. Elsevier Science, 2006.
[37] Giorgios Stoilos, Giorgios Stamou, Vassilis Tzouvaras, Jeff Z. Pan, Ian Horrocks. "Fuzzy OWL: Uncertainty and the Semantic Web". International Workshop of OWL: Experiences and Directions, vol. 21, no. 5, pp. 84-87, 2005.
[38] Jun Zhai, Wei Liu, Yiduo Liang, Jiatao Jiang , "Fuzzy knowledge representation for fuzzy systems based on ontology and RDF on the Semantic Web", International Conference on Information and Automation, pp. 1101 – 1105, 2008.
[39] Danijela Milosevic, Camil Sukic, Ramo Sendelj, "Ontology–based Learner Modeling in Intelligent Tutoring Systems", Technics Technologies Education Management, Vol. 5, no. 2, pp. 271-277, 2010.
[40] Jose Manuel Morales-del-Castillo, Eduardo Peis, Enrique Herrera-Viedma, "A Web-Based Fuzzy Linguistic Tool to Filter Information in a Biomedical Domain", 9th International Conference on Intelligent Systems Design and Applications, pp. 61-66, 2009.
[41] Antonio Martin, Carlos Leon. "Intelligent retrieval in a digital library using semantic web". IADAT Journal of Advanced Technology on Educaction, vol. 3, no. 3, pp. 427-429, 2009.
[42] Jun Zhai, Jianfeng Li, Yan Chen. "Knowledge modeling of product data based on fuzzy ontology". Applied Mechanics and Materials, Vol. 26-28, pp. 347-351. 2010.
[43] Leszek Borzemski, Krzysztof Zatwarnicki and Anna Zatwarnicka, "Adaptive and Intelligent Request Distribution for Content Delivery Networks", Cybernetics and Systems, vol. 38, no. 8, pp. 837-857, 2007.