A Fuzzy Ontology-Approach to improve Semantic Information Retrieval

Silvia Calegari[1] and Elie Sanchez[2]

[1] Dipartimento Di Informatica, Sistemistica e Comunicazione
Università di Milano – Bicocca
V.le Sarca 336/14, 20126 Milano (Italia)
calegari@disco.unimib.it

[2] LIF, Biomathematiques et Informatique Medicale
Faculte de Medecine (Universite Aix-Marseille II)
27 Bd Jean Moulin, 13385 Marseille Cedex5, (France)
elie.sanchez@medecine.univ-mrs.fr

Abstract. This paper shows how a Fuzzy Ontology based approach can improve semantic documents retrieval. After formally defining a Fuzzy Knowledge Base, it is discussed a special type of new non-taxonomic fuzzy relationships, called (semantic) correlations. These correlations, first assigned by experts, are updated after querying, or when a document has been inserted into a database. It is then introduced an Information Retrieval algorithm that allows to derive a unique path among the entities involved in the query in order to obtain maxima semantic associations in the knowledge domain.

1 Introduction: Fuzzy Ontology and Fuzzy Knowledge Base

Ontologies in the sense of a formal, explicit specification of a shared conceptualisation [1], constitute a key component of the Semantic Web, facilitating a machine processable representation of information. Two-valued-based logical methods are insufficient to handle ill-structured, uncertain or imprecise information encountered in real world knowledge. A tolerance for imprecision, by a positive use of Fuzzy Logic may be exploited to enhance the power of the Semantic Web [2, 3]. It has been shown that Fuzzy Logic allows to bridge the gap between human-understandable soft logic and machine-readable hard logic. Indeed there has been a natural integration of Fuzzy Logic in Ontology in order to define a new theoretical paradigm called Fuzzy Ontology [4, 5, 6].

Recently, an increasing number of approaches to Information Retrieval have proposed models based on concepts rather than on keywords. So that, in this work, ontologies have been combined to objects (stored in a database) in order to search new documents semantically correlated to user's query.

In this paper, the notion of Fuzzy Concept Network (FCN), introduced in [7], is extended incorporating Database Objects so that, concepts and documents can similarly be represented in the network. It is then introduced and described an Information Retrieval algorithm using an Object-Fuzzy Concept Network (O-FCN). This algorithm allows to derive a unique path among the entities involved in the query in order to obtain the maximum semantic associations in the knowledge domain.

It will now be introduced a formal Fuzzy Ontology (see also [4, 5]). This approach depends purely on an application choice. Indeed, we consider a formal Fuzzy Ontology as a quadruple $\mathbf{O}_{\mathcal{F}} = \{\mathbf{C}, \mathbf{R}, \mathbf{F}, \mathbf{A}\}$ where $\mathbf{C}$ is a set of fuzzy concepts, or entities indifferently. The set of entities of the fuzzy ontology will be indicated by $\mathbf{E}$. $\mathbf{R}$ is a set of fuzzy relations. Each $R \in \mathbf{R}$ is a n-ary fuzzy relation on the domain of entities $R : \mathbf{E}^n \mapsto [0, 1]$. In particular, $\mathbf{R} = \mathcal{T} \cup \mathcal{T}_{not}$ where $\mathcal{T}$ is the set of the taxonomic relations and $\mathcal{T}_{not}$ is the set of the non-taxonomic relations. $\mathbf{F}$ is a set of fuzzy relations on the set of entities $\mathbf{E}$ and a specific domain contained in $\mathcal{D} = \{integers, strings, ...\}$, and $\mathbf{A}$ is a set of axioms expressed in an proper logical language.

Note that even an OWL ontology "may" only include instances: we separated them in our approach, the advantage is that we can have one ontology and multiple instances that conform to it. Using this definition, it is possible to introduce the notion of Fuzzy Knowledge Base. Our definition is based on the vision of an ontology for the Semantic Web where knowledge is expressed in a Description Logic-based ontology as a triple $\langle \mathcal{T}, \mathcal{R}, \mathcal{A} \rangle$ where $\mathcal{T}$, $\mathcal{R}$ and $\mathcal{A}$ are respectively a TBox, RBox and ABox [8]. Thus, by using a fuzzy ontology the knowledge of a domain is defined in order to correspond to a Description Logic (DL) knowledge base.

Definition 1. *A Fuzzy Knowledge Base is a couple defined as:*

$$\mathbf{KB}_{\mathcal{F}} = (\mathbf{O}_{\mathcal{F}}, \mathcal{I})$$

where $\mathbf{O}_{\mathcal{F}}$ is a Fuzzy Ontology as previously defined and $\mathcal{I}$ is a set of instances associated with the fuzzy ontology. Furthermore, every concept $C \in \mathbf{C}$ is a fuzzy set on the domain of the instances defined as $C : \mathcal{I} \mapsto [0, 1]$.

In this context the set $\mathcal{I}$ is identified with the objects stored in the database, i.e. $\mathbf{O_{DB}} = \mathcal{I}$ and $C : \mathbf{O_{DB}} \mapsto [0, 1]$. In particular the set of objects can consist of documents, digital pictures, notes and so on, i.e. $\mathbf{O_{DB}} = \{\mathcal{D}, \mathcal{P}, \mathcal{N}, \dots\}$ where $\mathcal{D}$ is a set of documents, $\mathcal{P}$ is a set of digital pictures, and $\mathcal{N}$ is a set of notes, etc.

A new fuzzy relationship: Correlation. In the Semantic Web area of research, a crucial topic is to define a dynamic knowledge of a domain adapting itself to the context. *In order to achieve this aim, it is needed to handle the trade off between the correct definition of an object (given by the ontology structure) and the actual meaning assigned to the artifact by humans (i.e. the experience-based context assumed by every person according to his specific knowledge).*

In [7] it has been proposed a system that allows to achieve these objectives. It consists in the determination of a semantic correlation among the entities that are searched together, for example, in a query or when a document has been inserted into the database. In particular, a fuzzy weight on the correlations is also assigned during the definition of the ontological domain by an expert according to his/her experience. A *correlation* is a binary non-taxonomic fuzzy relation: $corr : \mathbf{E} \times \mathbf{E} \mapsto [0, 1]$, where $\mathbf{E} = \{e_1, e_2, \dots, e_n\}$ is the set of the entities contained in the ontology. This defines how the entities are linked semantically. The closer to 1 is the $corr$ value, the more the two considered entities are semantically associated.

In this way, the fuzzy ontology gives a solution to the trade off of the knowledge base and allows to dynamically adapt itself to the context in which it is introduced.

2 Information Retrieval Algorithm using O-FCN and its Evaluation

In [7] we introduced a Fuzzy Concept Network (FCN) to represent the dynamical behaviour of the fuzzy ontologies. In particular, the FCN representation lets us introduce a new semantic network based on the correlations defined in the fuzzy ontology. But an ontology allows to handle a complete knowledge base and so to make reasoning on the instances. In this work we extend this possibility by inserting directly in the FCN the objects of the domain stored into the database. In this way, we can reason directly with the elements of the specific application only visiting the FCN graph. In the following an extended FCN definition is given in order to insert the objects of the domain:

Definition 2. *An Object-Fuzzy Concept Network (O-FCN) is a weighted graph* $\mathcal{N}_{fo} = \{\mathbf{O_{DB}}, \mathcal{N}_\mathbf{f}\}$*, where* $\mathbf{O_{DB}}$ *is the set of the objects stored in the database and* $\mathcal{N}_f = \{\mathbf{E}, F, m\}$ *is a Fuzzy Concept Network (FCN). Each object is described by the entities of the FCN , i.e.* $\forall o_i \in \mathbf{O_{DB}}\ \ \mathbf{o_i} = \{\mathbf{e_1}, \ldots, \mathbf{e_n}\}$ *where* $\mathbf{e_i}, \ldots, \mathbf{e_n} \in \mathbf{E}$*.*

The set $\mathbf{O_{DB}}$ identifies all the information that is contained into the database, such as documents, digital pictures, videos, and so on.

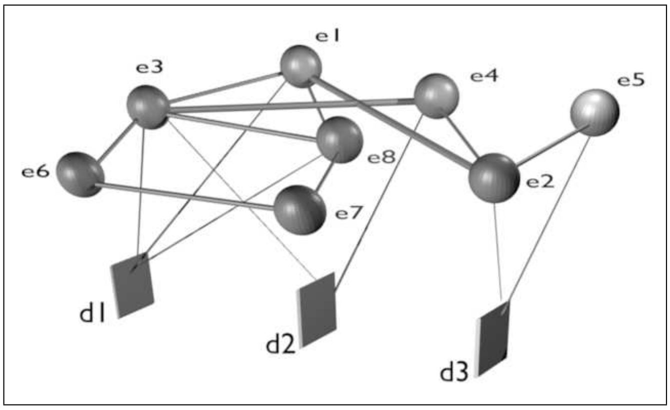

Fig. 1. A graphical representation of an Object-Fuzzy Concept Network.

In Fig. 1 it is given a 3D graphical representation of the prototype of a small O-FCN. The different thickness of the links identifies how strongly the entities are correlated. The thicker the link the more correlated are the two entities (i.e. the closer to 1 is the fuzzy value).

A recent application of Information Retrieval System (IRS) is the Semantic Web area of research. Indeed, the necessity of a better definition of IRS emerged in order to retrieve semantic information considered useful to a user query. Information Retrieval is a domain that involves the organization, storage, retrieval and display of information [9]. In order to extend the query vector it has been proposed a new algorithm based on fuzzy ontology. When navigating the O-FCN it is possible to find semantic links among the concepts: for each term specified in the query, a unique path is defined at each step, corresponding to the maximum value correlation. A step-by-step brief description of this new algorithm is given below (see also Fig. 2):

```
'O-FCN'-IR Search ( $E_q$ : word vector )
1:    'O-FCN'-based $E_q$ extension (pruning phase)
2:    'O-FCN'-based documents extraction
3:    'O-FCN'-based relevance calculation (cosine distance)
return ranking of the documents
```

Fig. 2. New Information Retrieval Algorithm using O-FCN

The O-FCN has been involved in all the steps of the algorithm in order to semantically enrich the results that were obtained. In this way, to retrieve documents it is easier to process than from the previous one that used only FCN [7]. The algorithm input is a vector E_q identifying the terms in the query. The first step (1) uses these terms to locate the unique path finding maximum correlation value among them. E_q is extended navigating the O-FCN recursively. Now, the “pruning phase” is directly inserted into the query extension algorithm. In this way, it is possible to find immediately the important entities, which are more semantically correlated w.r.t. the E_q set. In step (2) the O-FCN has been involved in order to directly extract the documents by the network. Whereas in the last step, O-FCN is used to calculate the relevance of the documents in order to sort them in decreasing order. The final score of a document is evaluated through a cosine distance among the weights of each entity. This is done for normalisation purposes. Such a value is finally sorted in order to obtain a ranking among the documents.

Evaluation A creative learning environment is the context chosen to test the new Information Retrieval algorithm based on O-FCN. In particular, the ATELIER (Architecture and Technologies for Inspirational Learning Environments) project has been involved. ATELIER is an EU-funded project that was part of the Disappearing Computer initiative. The aim of this project was to build a digitally enhanced environment, supporting a creative learning process in architecture and interaction design education. In this context, it emerges that the evolution of the O-FCN is mainly given by the words of the documents inserted in a hyper-media data base (HMDB) and from the entities written during the definition of a query by the students.

We have studied the dynamic evolution of the O-FCN examining 485 documents and 200 queries of the students. For each query a user had the opportunity to include up to 5 different concepts and the possibility to semantically enrich his/her requests using the following list of concept modifiers: *little, enough, moderately, quite, very, totally*.

The algorithm has been tested in two different situations: classical and fuzzy approaches. In the first case, the crisp situation has been reported assigning value 1.0 to the correlations values and without taking the concept modifiers into the queries of the students. Instead, in the last case, all the parameters described in this paper have been considered.

Fuzzy recall and fuzzy precision measures [10] are the parameters used in order to evaluate retrieval algorithms in these two different situations: crisp and fuzzy cases. In Table 1 it is reported the average values of fuzzy precision and fuzzy recall for the 200 queries performed in the two approaches. Retrieved documents are ranked up to a theta threshold (θ). In particular, we have chosen three values of θ ($0.35, 0.50$ and 0.75), to validate the algorithm in different situations.

Table 1. Average values of Fuzzy Precision and Fuzzy Recall in the fuzzy and crisp cases.

	Fuzzy Case		**Crisp Case**	
θ **value**	F. Precision	F. Recall	F. Precision	F. Recall
0.35	0.573	0.612	0.590	0.622
0.50	0.602	0.523	0.604	0.593
0.75	0.912	0.221	0.942	0.234

In Table 1, apparently, crisp approach is similar to the fuzzy one and the relevance of the obtained documents is more or less the same. Instead, in the fuzzy case it has been observed a better accuracy of relevant documents. Indeed, this result was derived from the analysis of *coefficient variance* based on fuzzy precision measure (here CV_P). In detail, $CV_P := (\frac{\sigma}{P_F}) \cdot 100$ where σ is the standard deviation calculated on the relevance of the documents and P_F is the fuzzy precision, and it is a useful statistic for comparing the degree of variation from one data series to another. In general, the larger this number, the greater the variability in the data. Figure 3 depicts the trend of CV_P between fuzzy and crisp approaches, for each query. In the fuzzy case we can observe higher CV_P values for the fuzzy case, for all the queries analysed. This means that the fuzzy case approach identifies more refinement and accuracy than the crisp case.

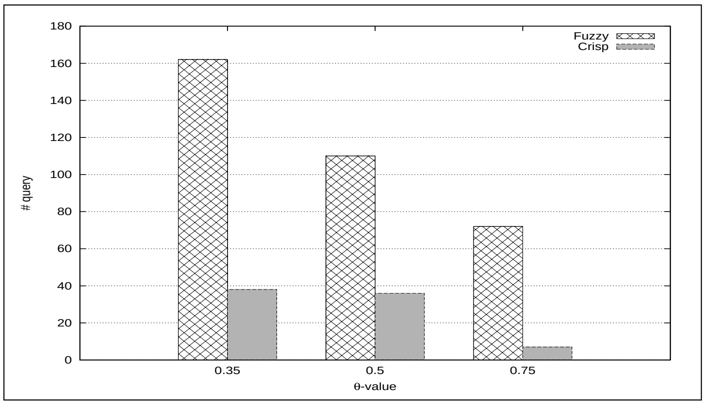

Fig. 3. Trend of CV_P value for each query.

3 Conclusion

It has been shown how the introduction of Fuzzy Ontologies, derived models and new structures, can improve an Information Retrieval System. More extensive developments will be shown in a forthcoming journal paper. The methodology allows to handle a trade off between the correct definition of an object, taken in the ontology structure, and the actual meaning assigned by individuals. So that it offers the opportunity to exploit an additional knowledge hidden in entities-documents relationships, or semantic correlations, after querying a database, but also to enrich the semantics of the system. After analysis, the obtained results for relevance presented a better accuracy in the fuzzy case than in the crisp one.

Acknowledgements

The work presented in this paper had been partially supported by the ATELIER project (IST-2001-33064). Particular thanks are due to Fabio Farina for his contribution in the Section of the numerical validations.

References

[1] Gruber, T.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition **5** (1993) 199–220

[2] Sanchez, E.: Fuzzy Logic and the Semantic Web. Capturing Intelligence. Elsevier (2006)

[3] Zadeh, L.: From Search Engines to Question-Answering Systems - The Problems of World Knowledge, Relevance, Deduction and Precisiation. In Sanchez, E., ed.: Fuzzy Logic and the Semantic Web. Capturing Intelligence. Elsevier (2006) 163–210

[4] Calegari, S., Ciucci, D.: Fuzzy Ontology, Fuzzy Description Logics and Fuzzy-OWL. In: Proceedings of WILF 2007. Volume 4578 of LNCS. (2007) In printing.

[5] Calegari, S., Ciucci, D.: Fuzzy Ontology and Fuzzy-OWL in the KAON Project. In: FUZZ-IEEE 2007. IEEE International Conference on Fuzzy Systems (2007) In printing.

[6] Sanchez, E., Yamanoi, T.: Fuzzy ontologies for the semantic web. In Larsen, H.L., Pasi, G., Arroyo, D.O., Andreasen, T., Christiansen, H., eds.: FQAS. LNCS 4027, Springer (2006) 691–699

[7] Calegari, S., Farina, F.: Fuzzy Ontologies and Scale-free Networks Analysis. International Journal of Computer Science and Applications **IV**(II) (2007) 125–144

[8] Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F., eds.: The Description Logic Handbook: Theory, Implementation, and Applications. In Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F., eds.: Description Logic Handbook, Cambridge University Press (2003)

[9] Salton, G., Mcgill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York, NY, USA (1986)

[10] Sanchez, E., Pierre, P.: Fuzzy Logic and Genetic Algorithms in Information Retrieval. In Yamakawa, T., ed.: Proceedings of the 3rd Int. Conf. on Fuzzy Logic, Neural Nets and Soft Computing, Jono Printing Co. (1994) 29–35