1

Making Large Information Sources Better Accessible Using Fuzzy Set Theory

Guy De Tré[1]

1.1 Introduction

Nowadays our society still witnesses an ever growing amount of digital information sources that are made publicly accessible via the internet. Along with the availability of the huge quantity of data comes the need for query engines and tools to efficiently explore and access these data and provide users with the facilities to retrieve exactly what they are looking for. As users most efficiently express their retrieval preferences using natural language and as matching in information retrieval and query processing in such cases often becomes a matter of degree or in some cases even a matter of uncertainty, fuzzy set theory and its related possibility theory offer an excellent mathematical basis for the development of advanced data access methods. These observations are the rationales behind the research of the Database, Document and Content Management group at Ghent University. In what follows we briefly describe the evolution of our research in the recent twenty years. Herewith we also try to give insight in the developments in fuzzy set theory and information management that formed the inspiration for this evolution. Furthermore, we present our vision on some trends for future developments.

1.2 Early personal research experiences

The first time I came in contact with fuzzy set theory was at the beginning of the 90's during a math course taught by Etienne Kerre. Etienne has just finished his book on 'Introduction to the Basic Principles of Fuzzy Set Theory and some of its Applications' and was talking with such an enthusiasm about its topics and their potential applications that without any doubts this theory must have been something really beautiful. At that time I could not even imagine how fuzzy sets would have an impact on my future work and life.

I was lucky to start my professional career as knowledge engineer in a small spinoff company of the Artificial Intelligence lab of the Vrije Universiteit Brussel. Here I had to study the problem of efficient time modelling for

[1] Ghent University, Dept. of Telecommunications and Information Processing, Sint-Pietersnieuwstraat 41, B-9000 Ghent (Belgium)

complex train schedules. After two years I received an opportunity from Rita De Caluwe to join the Computer Science Laboratory and start PhD studies at Ghent University. This was the start of my academic career and at the same time for me a re-initiation to fuzzy set theory. My first research topic was the definition of a flexible, fuzzy time model able to handle complex time indications, as often encountered in natural language expressions, and the incorporation of this time model in a temporal database framework.

My first scientific research results were presented at the EUFIT'97 conference in Aachen [3] where I also attended the plenary talk by Lotfi Zadeh on the usefulness of generalized constraints. I was impressed by Lotfi's talk, but even more impressed by his openness and willingness to briefly discuss some of my questions and comments during the coffee breaks. The concept of a generalized constraint started intriguing me and was later on the basis of my PhD work. In this work I studied the use of generalized constraints in fuzzy object-oriented database modelling [4]. Hereby, generalized constraints, as proposed by Zadeh, were used for semantic data integrity modelling purposes, as well as for query formulation purposes. Uncertainty in the stored data was modelled using possibility distributions as initially proposed in [10], whereas for the evaluation of generalized constraints, an multiple-valued possibilistic logic, based on possibilistic truth values [9], but extended to explicitly cope with missing information has been developed [5]. I obtained my PhD, entitled 'A formal generalized object oriented database model, appropriate for the exploitation of crisp and non-crisp information', in applied sciences at Ghent University in June 2000.

After obtaining my PhD, I continued specialising myself in fuzziness and soft computing in database management and information retrieval. Hereby, investigating among others, the application of generalized constraints for the modelling of fuzzy and uncertain spatio-temporal information in databases [6], the use of level-2 fuzzy sets for dealing with concurrent, orthogonal occurrences of fuzziness and uncertainty in fuzzy database modelling [7], and the handling of null values in fuzzy databases [8].

1.3 Research at the DDCM lab

In 2004, I obtained a tenured professor position at the Faculty of Engineering and Architecture of Ghent University with research area 'fuzzy information processing'. In that year we also established the Database, Document and Content Management (DDCM) research group by restructuring and renaming the former Computer Science Laboratory, now explicitly focussing its research, education and service activities on the handling and management of (imperfect) information.

The research mission of the DDCM group is to search for new soft computing techniques allowing to make large, heterogeneous data collections better accessible. The rationale behind this mission is in essence to find solutions for the demand of our society to handle the ever growing amount of digital information sources more efficiently. Additionally, the envisioned research offers better potentials for industrial and practical applications than pure fuzzy database modelling research offers, what is an important consideration in Ghent University's engineering faculty. The main research topics of the group include:

- **Coreference detection.** An important issue in database querying and information retrieval encompasses the task to find out whether two pieces of information refer to the same real world entity or not. If this is the case, we call the pieces coreferent. Beside being important for data access purposes, coreference detection is also useful to help guaranteeing data quality. Our group studies coreference detection of atomic data, collections, structured data, multimedia, texts and more general unstructured data. Among the applications under development is the ear identification application which is established in close cooperation with the Medical Imaging Center of the KULeuven and the Disaster Victim Identification team of the Belgian Federal Police. Coreference detection can be done both on the data (e.g., database) and on the metadata (e.g., database schema).
- **Information fusion.** Once detected, coreferent data can be further processed. For example, in database context, storage of coreferent data should be avoided as this would imply the storage of duplicated and often inconsistent information. As a solution the coreferent data can be merged or fused. The fusion of coreferent data is being studied by the group. Special research focus goes to the fusion of texts in the context of multiple document summarization.
- **Spatio-temporal information modelling.** Spatial and temporal data are generally recognised as special characteristics of information which deserve special care in database and information system contexts. Indeed, a lot of facts are registered in a database in a given spatial and or temporal context. Our research studies the handling of imperfect (imprecise, uncertain, incomplete) spatio-temporal information. For the handling of imperfect temporal information we closely cooperate with Olga Pons of the University of Granada. An application dealing with imperfect time indications in a database of medieval charters is under development. For the spatial data processing we cooperate with Nico Van de Weghe of the Geography department of Ghent University and with Jörg Verstraete who is currently employed at the Systems Research Institute of the Polish Academy of Sciences. Current research includes the efficient handling and analysis of moving objects, a research topic where also Bernard De Baets of Ghent University is involved in.
- **Bipolar information handling.** The DDCM research group also has an excellent cooperation with Sławomir Zadrożny and Janusz Kacprzyk of the Systems Research Institute of the Polish Academy of Sciences. This joint

research can be best categorised under fuzzy querying and fuzzy databases and comprises the handling of bipolarity in both database querying and database modelling. Bipolarity hereby refers to the fact that information, as communicated by humans, often has positive and negative components which do not necessarily have to complement each other. Extensions of fuzzy set theory, based on interval-valued fuzzy sets, Atanassov's intuitionistic fuzzy sets or twofold fuzzy sets form an excellent basis for further research on bipolarity.

Fig. 1.1. From left to right: Jozo Dujmović, Lotfi Zadeh and Guy De Tré at the WConSC'11 conference diner, San Francisco, 2011.

- **Decision support.** Multiple criteria decision support systems have many things in common with flexible querying systems. In both systems, the user has to specify criteria which have to be evaluated (for each case under consideration, resp. for each relevant database record) and in both systems evaluation results have to be aggregated to an overall degree of suitability or degree of satisfaction. The DDCM research group closely collaborates with Jozo Dujmović of San Francisco State University to study the applicability of the general logical scoring of preference (LSP) method for geographical suitability map construction and more recently for efficiently dealing with the opinions of multiple decision makers. Another research aspect concerns the further improvement of fuzzy querying techniques with LSP facilities.

1.4 Some future trends

There is currently a high demand from industry to manage the tremendous amount of unstructured data like texts as easy as structured database data.

This implies that there is a need for semantic richer text interpretation and text analysis algorithms. For that reason, we foresee in the near future a growing importance of semantic rich text parsing mechanisms which allow to extract essential information and context from texts. Such mechanisms would also allow for smarter indexing and information retrieval techniques and will hopefully bring us a step closer to automatic ontology generation. Protoforms, as proposed by Lotfi Zadeh, could play an important role in such developments. Beside textual information, the content-based retrieval of multimedia documents like photographs, audio and video is in our opinion another challenge for future research.

By considering texts as sequences of words which on their turn are sequences of characters it is worth to investigate whether symbol sequences obtained from the annotation of sensor data or biomedical data (DNA, proteins, peptides, etc.) could be meaningfully processed as texts as described above. This, in order to obtain a semantic richer interpretation of these data.

Further research in the above mentioned areas is required and planned by the DDCM research group. The future will learn us whether this research could bring us a little bit closer to tools for the efficient and full exploration of the huge, ever growing quantity of data that is and still becomes available through the internet and the information systems of organisations, companies, societies, etc. We at least are enthusiastic and well motivated to tackle these challenges.

References

1. B.P. Buckles, F.E. Petry, “A fuzzy representation of data for relational databases”, *Fuzzy Sets and Systems* 7:213–226, 1982.
2. R. Cavallo, M. Pittarelli, “The theory of probabilistic databases”, *Proceedings of the VLDB’87 Conference*, Brighton, England, September 1-4 1987, pp. 71-–81.
3. G. De Tré, R. De Caluwe, B. Van der Cruyssen, “Dealing with Time in Fuzzy and Uncertain Object-Oriented Database Models”, *Proceedings of the EUFIT’97 Conference*, Aachen, Germany, September 9-11 1997, pp. 1157–1161.
4. G. De Tré, R. De Caluwe, B. Van der Cruyssen, “A Generalised Object-Oriented Database Model”, in: G. Bordogna, G. Pasi (eds.), *Recent Issues on Fuzzy Databases*, Studies in Fuzziness and Soft Computing, Physica-Verlag, Springer, Heidelberg, Germany, 2000, pp. 155–182.
5. G. De Tré, “Extended Possibilistic Truth Values”, *International Journal of Intelligent Systems*, 17(4):427–446, 2002.
6. G. De Tré, R. De Caluwe, A. Hallez, J. Verstraete, “Modelling of Fuzzy and Uncertain Spatio-Temporal Information in Databases: A Constraint-based Approach”, in: B. Bouchon-Meunier, L. Foulloy, R.R. Yager (eds.), *Intelligent Systems for Information Processing: From Representation to Applications*, Elsevier Science B.V., Amsterdam, the Netherlands, 2003, pp. 117–128.
7. G. De Tré, R. De Caluwe, “Level-2 fuzzy sets and their usefulness in object-oriented database modelling”, *Fuzzy Sets and Systems*, 140(1):29–49, 2003.
8. G. De Tré, R. De Caluwe, H. Prade, “Null values revisited in prospect of data integration”, *Lecture Notes in Computer Science*, 3226:79–90, 2004.
9. H. Prade, “Possibility sets, fuzzy sets and their relation to Lukasiewicz logic”, *Proceedings of the 12th International Symposium on Multiple-Valued Logic*, Los Angeles, USA, 1982, pp. 223–227.
10. H. Prade, C. Testemale, “Generalizing Database Relational Algebra for the Treatment of Incomplete or Uncertain Information and Vague Queries”, *Information Sciences*, 34:115–143, 1984.