Creation and Enhancement of Fuzzy Ontology Structures for Query Answering from Text Documents – A Text Mining Paradigm

by

Muhammad Abulaish

Department of Mathematics

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the

Indian Institute of Technology Delhi

April 2006

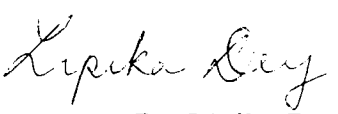
To my parents, who taught me how to tackle life's many impediments.

To my wife, sons – Yusuf and Haris, and daughter – Aisha who make the journey worthwhile.

Certificate

This is to certify that the thesis entitled **"Creation and Enhancement of Fuzzy Ontology Structures for Query Answering from Text Documents – a Text Mining Paradigm"** submitted by Mr. Muhammad Abulaish to the Indian Institute of Technology Delhi, for the award of the degree of Doctor of Philosophy, is a record of the original bona fide research work carried out by him under my supervision and guidance. The thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree or diploma.



New Delhi
April 2006

Dr. Lipika Dey
Supervisor

Acknowledgements

I am very grateful to my supervisor, Dr. Lipika Dey, for having accepted me as a student and introducing me to such an interesting field. With her sagacious suggestions, understanding, patience, enthusiasm, responsibility, vivid discussions, insight, sensibility, observation, rigor, and above all with her appreciating nature, she helped me to work in a right perspective. I cannot thank her enough for the time and effort she spent in disciplining my style of working, introducing me to new problems and helping me put things in right perspective, professionally and personally. I am deeply indebted to her for her intense involvement, effort, advice, encouragement, patience and help. Her deep insight into problems, imaginative ideas, zest for work, foresight, rigour and tremendous capacity for disciplined work, have been a source of inspiration and motivation for me. Working with her has been a day-to-day learning and invigorating experience. Inspite of being extremely busy, she could always find time for my work; means so much to me. Without her planning, immaculate work scheduling, active interest and positive thinking, my thesis would not have been through.

I take this opportunity to thank Prof. B. R. Handa who selected me and gave me a chance to get enrolled in the department for the Ph. D. program. I express my regards to Prof. B. Chandra, DRC Chairperson and Head of the Department, for her support. Special thanks to my SRC Members: Prof. Suresh Chandra, Dr. Nesar Ahmad and Dr Tapan Roy Chowdhury for spending their valuable time during the discussions over seminars. I am grateful to my pre-Ph.D. course instructors: Prof. B. Chandra, Dr. Lipika Dey, Dr. Wagesh Shukla, Dr. Niladri Chatterjee and Dr. S. Chandra Shekhar Rao for their valuable suggestions as well. Further, I extend my thanks to Dr. R. K. Sharma, Dr. S. Dharmaraja and other faculty members of the department for their encouragement. I am also thankful to Mrs. Rashmi

Wadhwa and other non-teaching staff members of the department for all their help and good wishes. I am also thankful to I.I.T. Delhi authorities for providing me the necessary facilities.

It is a pleasure for me to thank Dr. Shailendra singh and Dr. K. V. Krishna for their suggestions to get acquainted with the department. I am also thankful to Mr. Sachin Kumar and all other research scholars of the department for their company.

I am very grateful to the Department of Science and Technology, Ministry of Science and Technology, who sponsored me to attend the CIMPA-UNESCO-INDIA school on "Soft Computing Approach to Pattern Recognition and Image Processing", held at Indian Statistical Institute, Kolkata during December 2-13, 2002 which was the foundation on which I started my research work.

I would also like to thank Dr. Siraj Hussain (IAS), ex Vice-Chancellor of Jamia Hamdard (Deemed to be University) for his encouragement and permission to get enrolled for the Ph.D. program.

I am also thankful to my elders and colleagues, especially Prof. Sharfuddin Ahmad, Prof. M. R. Khan, Dr. Naseem Ahmad, Dr. Shehzad Hasan and other staff members of my working department for their kind support without which it would not be possible for me to finish this work on time. I owe special thanks to my friend Mr. Zahiruddin for his kind support.

To my family, I owe so much more than words can express, for what they have done for me. It was only because of them that I could do a Ph. D. Their constant encouragement, love, support and advice have been so important. For being there for me at all times and for taking over all the problems so that I could concentrate single-mindedly on my research.

Finally, I would like to thank almighty God who helped me at different critical stages during this work and shown me the right way when I was in trouble.

Md. Abulaish

New Delhi

Muhammad Abulaish

Abstract

Over the last decade the World Wide Web (WWW) has virtually become the main repository for data sharing and information gathering. Most of the data on the Web are embedded within text documents which are either unstructured or semi-structured in nature. Thus even though there is no scarcity of information on the Web, locating, extracting and analyzing required information from this vast unstructured collection is a complex and challenging task. Most of the problems are obviously due to the incapability of the computer in comprehending natural language texts with all its nuances.

Word based searches for relevant information from texts retrieve a huge collection and burden the user with information overload. Ontology based text information retrieval can perform concept-based search and extract only relevant portions of text containing concepts that are present in the query or those that are semantically linked to query concepts. While these systems have better precision of retrieval than general-purpose search engines, there are some crucial problems which have to be tackled while designing such systems. Considering that ontologies are created and maintained by domain experts, creation and maintenance of ontologies are expensive tasks. Though ontology provides a structured framework to store domain knowledge, all ontological concepts cannot be unambiguously described using crisp property descriptors. Besides, the ontological descriptors may not exactly match text descriptions or the user given descriptors in query. Appropriate reasoning mechanisms have to be designed in order to apply ontologies for text information processing.

In this thesis we have proposed the use of a text mining framework which can integrate the problems of ontology enhancement with text information retrieval. Thus the system provides

dual advantage of content-based retrieval and ontology learning from text documents. In order to achieve flexibility in concept definition we propose to enhance the traditional ontology structures into fuzzy ontology structures. These structures can accommodate imprecise concept descriptors that are prevalent in text documents. The ontology-based text mining framework mines precise and imprecise concept descriptions from text documents and enhances the fuzzy ontology structure with them. The mined knowledge is stored in a structured knowledge base which design is derived from the underlying fuzzy ontology structure representing the domain. We have shown that the fuzzy ontology structure can provide a viable solution to resolve inconsistencies over concept descriptions encountered in multiple overlapping ontologies. The proposed framework is also enriched with a fuzzy reasoning mechanism that is used to process user queries over a curated knowledge base.

The text-mining framework is enhanced to mine inter-concept relations besides structural semantic relations from text. The query processing principle is then appropriately extended to answer relation-based complex user queries over a curated database. This has been tested over biomedical document repositories, in which relations play an important role in focused information retrieval. Since mining can yield a wide array of relations, these relations are subjected to a significance analysis to identify the key relations of a domain. The significant relations are used for enhancing the underlying concept ontology into a relational ontology.

Table of Contents