Rough Fuzzy Approach for User-Centric Text Information Retrieval

by

Shailendra Singh

Department of Mathematics

Submitted

in fulfillment of the requirements of

the degree of

Doctor of Philosophy

to the

Indian Institute of Technology, Delhi

Hauz-Khas, New Delhi -110016, INDIA

January, 2004

Certificate

This is to certify that the thesis entitled **"Rough Fuzzy Approach for User-Centric Text Information Retrieval"**, which is being submitted by Mr. Shailendra Singh to the Department of Mathematics, Indian Institute of Technology, Delhi, for the award of the degree of Doctor of Philosophy, is a record of bonafide research work carried out by him under my guidance and supervision. 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 to any other university or institute for the award of any degree or diploma.

Dr. Lipika Dey

Assistant Professor

Department of Mathematics

Indian Institute of Technology, Delhi

Delhi (INDIA)

Dedicated

to

my father, mother

&

brother

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. 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 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, that she could always find time for my work; means so much to me. But for her planning, immaculate work scheduling, active interest and positive thinking, my thesis would not have been through.

I sincerely thank all the Faculty members of the Department of Mathematics, for being so supportive and helpful. I wish to thank Prof. Suresh Chandra for his time-to-time suggestions and motivations. I take this opportunity to thank Prof. J.B.Srivastava for all that he has done for me as the Head of Department and for his advice towards planning my future.

I am thankful to department staff members, for all their help and good wishes. Having such wonderful and cooperative staff in the Department made so many problems disappear. The complete freedom which I was allowed in the department computing laboratory has helped me a lot.

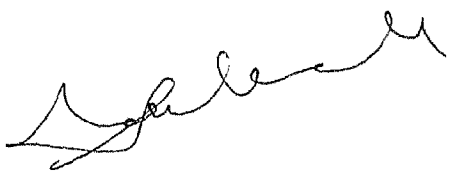
I am very grateful to Prof. Andrzej Skowron who had sent me the rough set tutorial which was the foundation on which I started my research work. He has also been a source of constant encouragement to me.

I would also like to thank Prof. Donald Kraft and Prof. Witold Pedrycz for helping me in more than one way. This work could not be completed without their support.

I am very thankful to all department research scholars with whom I have spent this long period and shared experiences. I want to give special thanks for support and help of all my friends, with whom I have enjoyed most of the moments of my stay in IIT.

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.

At last, I am very grateful to almighty GOD who has shown me the way when I was in trouble. I believe in GOD and did my work very confidently. I never forgot the proverbs: "Believe in yourself and do best", "Work is Worship", "Where there is will, there is way", and "Nothing is impossible in the world". These are the most important philosophy of my life.



Shailendra Singh

Abstract

In this thesis, we have considered the problem of user-centric text information retrieval from web documents. User relevance-feedback based methods are the most acceptable way to improve the precision of relevant information retrieval.

The problem of locating and retrieving relevant text documents is difficult since web documents are unstructured. We have proposed a complete framework to show how rough set based reasoning can be used to build a user relevance feedback based text filtering system, using concept approximations. The system uses discernibilty analysis to analyze the user's relevance feedback on a set of documents obtained with the initial query to obtain a more focused, modified query. We show that the precision of retrieval increases considerably with the use of the modified query.

The results of discernibility analysis are maintained by the system as the user's long-term interest representation that can be used to automatically scourge the web for new resources in the domain and present only the relevant ones to the user proactively. In order to provide customized information to the user, the documents fetched by the modified query have to be ranked by the system and only the relevant ones should be presented to the user. We have proposed rough-membership functions for computing document relevance in the concept approximation space. We have shown the effectiveness of the grading scheme through experimental results.

We have also integrated fuzzy reasoning concepts into our analysis to deal with certain aspects of natural language nuances like synonymy and provide fuzzy grading to documents. We have proposed a rough-fuzzy grading scheme that provides a visualization about the

quality of a document to the user. This unique visualization scheme allows the user to have an idea about the goodness or badness of a document.

Finally, we have tried to provide a framework whereby ontological structures can be used for concept approximations using rough sets. This can substantially improve the quality of knowledge retrieved within any particular domain. We have explored the possibilities of applying the granular approach of rough-set theory to model the inherent uncertainties in the domain of text information retrieval. Rough set based concept approximation techniques can be used very effectively to include related concepts from the ontology into a query. We have also shown how hedge-algebra based reasoning can be used to reason about properties of concepts. Since document descriptions are not likely to match ontological descriptions exactly, the approximations can pave the way for more intelligent retrieval.

Contents