

ON MEASUREMENT AND ENHANCEMENT OF WEB SEARCH QUALITY

by

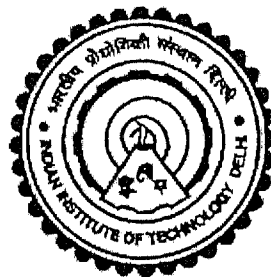
Mirza Mohd. Sufyan Beg

Electrical Engineering Department

Submitted

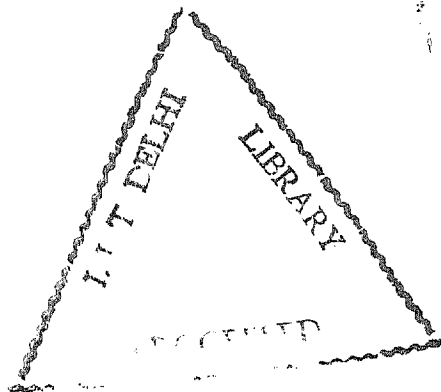
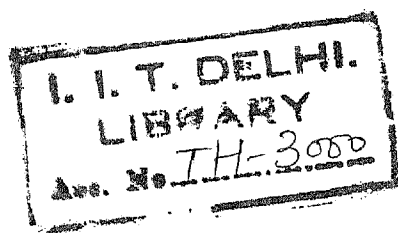
in fulfillment of the requirements of the degree of Doctor of Philosophy

to the



Indian Institute of Technology, Delhi

September 2002



TH
681.3 WWW
SUF-0

62
World wide web ;
search engines

Certificate

This is to certify that the thesis entitled "*On Measurement and Enhancement of Web Search Quality*" being submitted by Mirza Mohd. Sufyan Beg for the award of the degree of the **Doctor of Philosophy** to the Indian Institute of Technology, Delhi, is a record of the bonafide research work he has carried out under my supervision. The results contained in this thesis have not been submitted to any other University or Institute for the award of a degree or diploma.



(Dr. Nesar Ahmad)
Assistant Professor
Department of Electrical Engineering
Indian Institute of Technology, Delhi
New Delhi – 110 016
India

Acknowledgement

- AICTE — for providing me with the QIP fellowship
- AMU Aligarh — for sponsoring my Ph.D. program
- IIT Delhi — for being my seat-of-learning
- Dr. C. P. Ravikumar — for teaching me how to carry out research while he supervised me initially
- Dr. Nesar Ahmad — for adopting me when I was all alone and supervising me to the finish, so very considerately and affectionately
- Prof. H. M. Gupta
- Prof. Suresh Chandra
- Dr. Subrat Kar — for being on my research committee
- Prof. S. Prasad
- Prof. J. K. Chatterjee
- Prof. S. D. Joshi
- Mr. S. M. K. Rahman — for their constant encouragement
- Prof. L. A. Zadeh
- Dr. M. Nikraves
- Dr. M. J. Zaki
- Dr. Shaz Qadeer — for their useful comments
- Prof. S. A. Abbasi
- Prof. A. Mahanta
- Prof. M. Q. Rafiq — for being my mentors
- Sureshji
- Tiwariji — for their technical help
- Rakeshji
- Jileyji — for their non-technical help
- Mr. Rafat Jamal — for his ever available help
- Dr. Vineet Sahula
- Abhinav & Mona
- Abu Sarwat
- Mukesh
- Alok
- Gaurav — for their healthy discussions as fellow research scholars
- Anubhav
- Ranjan
- Manorajan
- NC-2 community — for being my friends on campus
- My parents — for what I am today
- My wife — for her constant support
- My In-laws — for “generously” providing me with that

My little daughters — "source" of constant support
— for their best effort to keep me away from
my research
My sister and her husband — for their own worries
My uncles and aunts — for their hospitality
My PC — for qualifying to be an object of jealousy for
my wife


(Mirza Mohd. Sufyan Beg)

Abstract

Search engines are among the most popular as well as useful services on the web. Many search engines are publicly available to Internet users today and more are likely to appear in the future. We propose a procedure for quantifying the quality of web search results obtained from a given search engine. We infer the feedback of the web searcher from his actions on those search results, like for instance, the sequence in which the user clicks the results, how much time he spends at those results, whether or not he prints, saves, bookmarks, e-mails or copies-and-pastes a portion of those documents. The document importance weight is then calculated either by a proposed expression, or alternatively by using Fuzzy Theory. Consequently, sorted in a descending order of their importance, the sequence of documents is compared with the original ranking given by a search engine. The resulting correlation coefficient, when averaged over a set of test queries, gives the required measure of search quality. We show our results for 7 public search engines with 15 arbitrarily taken queries. Our emphasis is more to demonstrate the procedure of search quality measurement than to carry out the actual performance measurement of these search engines.

The user feedback obtained for the purpose of measuring search quality, has also been utilized for improving upon the results. A user feedback based search engine is proposed, which increases the index weights of those documents that are being consistently preferred by the users, and decreases for those that are being left untouched by the users. We also propose a couple of novel additions to the web search query language, namely, the queries with its terms connected using the linguistic operators and those with the connectors being weighted logical operators. In the performance analysis experiment, we have used a probabilistic model to generate the user responses automatically. We observe that after sufficient learning has gone into our search engine, it renders a better level of user satisfaction.

The results obtained from different search engines have been aggregated to combine their individual advantages. For this problem of rank aggregation, the Genetic Algorithm (GA) has been first made use of. Also adopted are a couple of fuzzy rank ordering techniques like the one due to Shimura and the other one due to Dubois and Prade. But their poor performance when compared to the classical Borda's method of rank aggregation prompted us

to suggest some improvement in the Shimura technique based on the Ordered Weighted Averaging (OWA) operators. Two techniques, namely Membership Function Ordering (MFO) and Mean-By-Variance (MBV) are also put forward. The best performance, however, was found when a technique based on entropy minimization principle was investigated.

A notion of Biased Rank Aggregation is also proposed, wherein the rank aggregation is done with the importance of the result scaled by the search quality measure of the search engine returning that result. As an application of these works, namely, the search quality measurement, the user feedback based enhancement in search quality, and the biased rank aggregation of search results, we propose an Enhanced Web Search System. The proposed system maintains the profile of each individual user, and when queried, gives them the results suited to their taste, and conforming to their search history.

In memory of
my grandparents

Contents

List of Figures	xi
List of Tables	xiii
1 Introduction	1
1-1 Motivation	1
1-2 Challenges	2
1-3 Limitations in Existing Work	4
1-4 Open Problems	5
1-5 Outline of the Thesis	6
1-6 Contributions of this Thesis	8
2 Content-Based Web Search	10
2-1 Text Pre-Processing	11
2-2 Term Frequency	12
2-3 Vector Space Model	13
2-4 Boolean Similarity Measures	14
2-5 Enhanced Boolean Similarity Measure	17
2-6 Some Miscellaneous Works	20
2-7 Problems with Content Based Web Search	24
3 Connectivity-Based Web Search	26
3-1 Hyperlink Based Web Search	27
3-1.1 Integrating Content Based and Connectivity Based Techniques	34
3-2 Finding Related Pages	35
3-3 Web Characterization	38
3-4 Summary	42

4	Subjective Measurement of Web Search Quality	43
4-1	Introduction	43
4-2	Search Quality	44
4-2.1	User Feedback Vector	45
4-2.2	The Search Quality Measure (<i>SQM</i>)	47
4-2.3	Some Finer Points in SQM	50
4-3	Fuzzy Search Quality Measure (FSQM)	52
4-3.1	Step I of FSQM: Transformation of Each Component of User Feedback Vector to a Corresponding Fuzzy Preference Relation	54
4-3.2	Ordered Weighted Averaging (OWA) operators	55
4-3.3	Relative Fuzzy Linguistic Quantifier	56
4-3.4	Step II of FSQM: Aggregation of Individual Fuzzy Preference Relations	57
4-3.5	Step III of FSQM: From Combined Fuzzy Preference Relation to Combined Fuzzy Preference Ordering	58
4-3.6	Step IV of FSQM: Finally Getting FSQM	59
4-4	Experiments and Results	60
4-5	Discussion	66
4-5.1	About the Set of Test Queries	67
4-6	Summary	69
5	User Feedback Based Enhancement in Web Search Quality	70
5-1	Introduction	70
5-2	Search Engine Model	71
5-2.1	Query Handler	73
5-2.2	Query Formulation using Linguistic Operators	74
5-2.3	Query Formulation using Weighted Logical Operators AND, OR and NOT	76
5-3	Updating Relevance Matrix from the User Feedback	78
5-3.1	Dealing with Simple One-Keyword Query	78

5-3.2	Dealing with Complex Multiple Keywords Query	80
5-3.3	Dealing with NOT Terms in the Query	81
5-3.4	Dealing with Linguistic and Weighted Queries	82
5-3.5	Exponential Penalization for Documents Left Untouched by the User	83
5-4	Experiments and Results	84
5-4.1	Advantages of User Feedback Based Web Search	89
5-5	Summary	89
6	Rank Aggregation Based Enhancement in Web Search Quality	91
6-1	Introduction	91
6-2	Background and Related Work	93
6-3	Genetic Algorithm for Rank Aggregation	97
6-3.1	Objective Function for GA	97
6-3.2	Representation of Chromosomes in Decimal	98
6-3.3	Crossover	99
6-3.4	Mutation	100
6-4	Some other Soft Techniques for Rank Aggregation	101
6-4.1	Shimura Technique of Fuzzy Ordering for Rank Aggregation	101
6-4.2	Mean Position Based Rank Aggregation	102
6-4.3	Technique of Dubois and Prade (DP) for fuzzy ordering	103
6-4.4	Membership Function Ordering (MFO) Technique	105
6-4.5	Mean-By-Variance (MBV) Technique	107
6-4.6	Modified Shimura technique	108
6-4.7	Entropy Minimization Technique	109
6-5	Biased Rank Aggregation	111
6-6	Experiments and Results	113
6-6.1	Results of GA Based Rank Aggregation	114
6-6.2	Results of Other Soft Computing Techniques Based Rank Aggregation	117

6-7 Summary	118
7 Improvement and Integration	122
7-1 Improved Search Quality Measure	122
7-2 Enhanced Web Search System	128
7-2.1 Personal Web Searching	129
7-3 Experiments and Results for Enhanced Web Search System	132
7-3.1 Anticipated Advantages with the Enhanced Web Search System	133
8 Conclusion and Future Directions	135
8-1 Future Work	138
References	140
Bio-Data	148