ON MEASUREMENT AND ENHANCEMENT OF WEB SEARCH QUALITY

by

Mirza Mohd. Sufyan Beg

Electrical Engineering Department

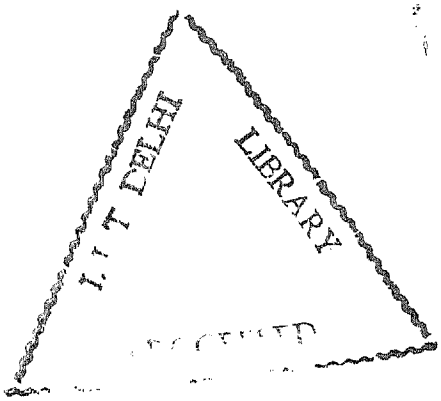
Submitted

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

to the

Indian Institute of Technology, Delhi

September 2002



By

World wide web ;
Search engines

Certificate

This it 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. Nikravesh
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

"source" of constant support

My little daughters — 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

M. M. J.

(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