A

Dissertation

On

Improved Rank Aggregation

For the submission of partial fulfillment of degree of

Master of Technology
In
Computer Science and Engineering
2014

Submitted by:
Mohd. Zeeshan Ansari
2K11/CSE/23

Under the supervision of:
Mr. Manoj Kumar
Associate Professor, DTU.

Department of Computer Engineering

Delhi Technological University

(Formerly Delhi College of Engineering)

New Delhi

Acknowledgement

I would like to express my deepest gratitude to my supervisor Mr. Manoj Kumar for his invaluable guidance and support. His creativity and innovative skills have always been a constant source of motivation for me. He is a great person and one of the best mentors, I always be thankful to him. I would also like to thank Prof. M. M. Sufyan Beg for devoting precious time his time in discussing ideas with me, teaching relevant topics and giving his invaluable feedback. I would like to extend the thanks to all the faculty members for their motivation and moral support. I would like to thank all my classmates for their cooperation and support. I would like to dedicate this dissertation thesis to my supportive parents and caring wife who have always been with me to make it a complete success.

Mohd Zeeshan Ansari
2K11/CSE/23,
Master of Technology,
In Computer Science & Engineering,
Delhi Technological University.

Certificate

This is to certify that the work submitted in this dissertation entitled **"Improved Rank Aggregation"** submitted in partial fulfillment for the award of degree of **Master of Technology in Computer Science & Engineering** at **Department of Computer Engineering, Delhi Technological University** by Mohd. Zeeshan Ansari, Roll No. 2K11/CSE/23 is carried out by him under my supervision and guidance. The matter embodied in this dissertation work has not been submitter earlier for the award of any degree or diploma in any institution to the best of my knowledge and belief. He has completed his work with utmost diligence and sincerity.

Mr. Manoj Kumar
Associate Professor,
Department of Computer Engineering,
Delhi Technological University.

Mohd. Zeeshan Ansari
2K11/CSE/23,
Master of Technology,
Computer Science & Engineering

Contents

List of Figures

List of Tables

Abstract

Multiple rankings of same object based on multiple criteria poses a problem of choice to rank that object at a position closest to all the rankings. Choosing a ranking for a list of such objects ranked previously is called rank aggregation. The aggregated ranking is analysed by computing the Kendall tau distance or Spearman Footrule distance. The ranking chosen by minimizing Spearman Footrule distance, is NP-Hard even if number of ranking lists greater than four for partial lists. In context of World Wide Web, the results generated by the multiple search engines may be collected together and an aggregated result can be produced using rank aggregation methods or in simple words the meta-search engine comes into existence. However, the use of existing search engines reveal that none of them have been effective in generating the results up to the quality and reliability desired by the end user, the reason being many.

The application of fuzzy logic techniques to minimize the aggregated Spearman Footrule distance obtained from rank aggregation when applied in meta-searching have been studied at length, and the improvements in the existing Shimura's technique, viz. Enhanced Shimura-square and Enhanced Shimura-sqrt have been proposed to achieve the extent that is better to Borda's Method, a benchmark for comparison of aggregated rankings as well the other common techniques. In addition, the improvements in existing heuristics based positional methods have also been proposed. A new heuristics have also been put forward for future work.

A series of experiments on real and benchmark data have been conducted to validate the proposed improvements. The platform used for experimentation was matlab framework. A comparative analysis of the previous and the proposed rank aggregation methods on six search engines and twenty one search queries have been presented. The different aggregation methods have been compared on the basis of aggregate spearman footrule distances. The computation time have also been measured and reported. It has been revealed by the experiments that our proposed soft computing method, Enhanced Shimura Square performs best in terms of effective performance and computational efficiency.

Keywords: Web, rank aggregation, meta-searching, spearman footrule distance, fuzzy logic.

Chapter 1

Introduction

1.1 Motivation

The fact that the hundreds of search engines are being available on the World Wide Web as of now, only one search engine is not sufficient in terms of effective results produced as well as the results of any one search engine are not reliable. Also the observation is that not a single search algorithm is widely acceptable and non is sufficient to span its coverage all over the web [1,6]. The search engines also suffer lot of drawbacks, firstly the indexing web data is time and space consuming. Due to rapid changes in the web data the search engine has to always manage a trade-off between the coverage, i.e., the number of documents that can be indexed with respect to entire web, and update frequency, i.e., the time elapsed between two successive procedures of re-indexing the complete database [4]. Secondly, the advertisers persuade a paid ranking that leads to loss of fair information and accuracy in rankings, popularly called as spammed pages in the context of search engines results [1,4]. Thirdly, the overall coverage also do not include the indexing of proprietary information available in online digital libraries, etc. [4]. Apart from the above mentioned demerits, a lot more have been written in literature by the critics.

1.2 Rank Aggregation

Rank aggregation problem has been in focus since long back finding its applications in social choice, where rankings given by multiple judges, each having a unique ranking criteria are combined together in order to achieve a **"single and combined"** ranking effect reflecting the opinion of all the judges.

Thus, given a set of individual rankings, each based on a unique ranking criteria, arriving at a consensus ranking that is closest to all the given rankings is called **rank aggregation problem**.

The rank aggregation problem also called rank fusion had been observed by different perspective in terms of permutations. With each permutation being a ranking of n elements, to compute a permutation that optimizes the aggregate distance between itself and each permutation [4].

The rank aggregation have been applied in social choice theory for voting [Borda 1781, Condorcet 1785]. In modern day applications, rank aggregation has been found to be extensively applied in information retrieval especially in meta-searching. The application into web meta-searching had been first studied in 2001 [1,8]. The other applications of rank aggregation include sports and competitions, collaborative filtering, database middleware, consumer ratings, etc.[9]

The performance of the rank aggregation techniques can be measured by the two most common distance measures, namely, Kendall tau distance and Spearman footrule distance. The distance measure of the final aggregated list is calculated with each of the input list. Subsequently, a normalized aggregated distance is then obtained. The optimization of this normalized aggregated distance is called **Optimal Aggregation Problem** which is NP-Hard in case of partial lists even when n>=4[1].

1.3 Rank Aggregation and Meta-Searching

Meta-searching, firstly, exploits the various techniques and algorithms of all the search engines by combining the results altogether. Secondly, it covers the entire World Wide Web, since various search engines have differing coverage area depending upon the underlying algorithm. Thus, by applying rank aggregation a lot of important achievements can be obtained, viz., if any search engine is giving a biased ranking to a document, the document is filtered out since the rest of the search engines give it a fair rank or do not rank it all in their rankings, thus enhancing the *spam reduction*. Further, the user can be relieved from the burden of search query formulation thus improving search through *word associations* [1,2].

The meta-search engines have gained importance over the single search engine due to *(i)*

increase in the coverage of web search space. The overlap being very small as 3% of the all the retrieved results, whereas unique results can be as high as 85% of the total results [6], *(ii)* meta-searching exploits multiple search engines thus enabling the consistency check on the results [6,8].

A lot of rank aggregation techniques have been found in literature which can be categorized into position based [2,10], order based [1,2], score based [2,10], heuristic based [2], learning based [11], probabilistic models [3,7], approximation algorithms [13], etc. The technique which rely only on the order of the search engines results from each search engine is called order based rank aggregation. The Borda's method [10] is the most popular and time efficient technique in this category. The mean rank aggregation [2] is another order based technique which is proved equivalent to the Borda's method. The techniques which assign a score, computed with the help of ranks, to each unique document are called score based methods. The Borda's method assigns a Borda Score to each document.

The rank aggregation function which employs a heuristic in the determination of rankings are heuristics based methods. The variance by mean, mean by variance, membership function ordering and entropy minimization [2] are the heuristics based techniques found in literature. The learning based methods include two categories supervised learning and unsupervised learning [11].

The approximation algorithms have also been employed in rank aggregation. Polynomial time approximation scheme (PTAS) which executes in doubly exponential in precision parameter [3], and 11/7-approximation algorithm [13] are approximation algorithms found in literature.

An approach to rank aggregation based on probabilistic models on rank permutation have been applied in past and the two most popular models among them are, Luce model and Mallows model. The Luce model decomposes the process of generating a permutation n objects into n sequential stages. At the k^{th} stage, an objects is selected and placed at position k, according to the probability based on the scored of unassigned objects. Mallows model is based on a distance model, which defines probability of a permutation based on its distance to a location permutation. Luce model has a linear complexity whereas Mallows model has a high computational complexity. Further a coset-permutation distance based stage-wise model with high efficiency have been proposed in [7].

Fuzzy logic techniques, viz. Shimura's technique, Dubois and Prade [7] have also been applied successfully in aggregation but with unsatisfactory results. Shimura's technique have been modified as improved Shimura which showed comparable results [5]. Membership function ordering have been employed which uses the same Gaussian membership function used by Dubois and Prade. We have implemented the Membership Function Ordering Technique and compared the results to our proposed techniques.

Chapter 2

Background

2.1 Definitions

We present here the definition used in the succeeding chapters.

2.1.1 Ordered Lists

Given a Universe U, an ordered list L = $\{l_1 >> l_2 >> l_3 >> \ldots >> l_{|M|}\}$ with each $l_i \in M \subseteq U$, and > is some ordering relation on M such that l_i is preferred over l_{i+1}. Let l_i denote the ranking of i^{th} element in L. A lower numbered position in a list denoted a higher rank. On assigning a unique identifier to each element in U, we may have, without the loss of generality, U = {1, 2, 3,…, |U|}.

The ranking are always taken positive, the topmost ranking given by a search engine is 1, followed by the higher rankings 2, 3, 4,… denoting the lower preferences of document in the list.

2.1.2 Full Lists or Complete List

If a list contains all the elements of U, then it is said to be a full list or complete list. That is, L contains all the elements of U or |L| = |U|.

Example 1: Given a full list L = {3,5,2,1,4} has the ordering relation 3>>5>>2>>1>>4, we have $l_1 = 3$, $l_2 = 5$, $l_3 = 2$, $l_4 = 1$ and $l_5 = 4$. It follows that U = {1, 2, 3, 4, 5}.

2.1.3 Partial Lists

If a list contains the elements which are only a subset of U, i.e., we have the strict inequality |L| < |U|, then it is said to be partial list.

Example 2: Given a lists L1 = {3, 5, 7, 6, 1, 4} and L2 = {2, 5, 3, 7, 1}. It follows that U = {1, 2, 3, 4, 5, 6, 7}. We have |L1| = 6 < |U| = 7 and |L2| = 5 < |U| = 7. The lists L1 and L2 are partial lists.

2.1.4 Top d lists

If a list L contains the top d elements of the subset M, where the elements not present in L are assumed to be ranked below d, is said to be a top d list, where d is size of list. That is, |L| = d. It follows that each element of L is ranked above all unranked elements.

Example 2: Given L1 = {3,5,2,1,4},

L2 = {2,5,3,7,6},

L3 = {5,3,2,4,8},

L4 = {2,5,3,6,7} and

L5 = {1,5,2,8,4}

We have U = {1,2,3,4,5,6,7,8} with |L1|=|L2|=|L3|=|L4|=|L5| = 5 < |U| = 8.

2.1.5 Random Lists

The lists created by using pseudo random number generation will be called as random lists through the text.

2.2 Distance Measures

The two leading distance measures found in literature are Kendall tau distance and Spearman footrule distance.

2.2.1 Kendall tau distance

Given two full lists L1 and L2, each of cardinality |L|, the Kendall tau distance is given by

$$K(L1, L2) = \frac{|\{(i,j) \mid \forall\, L1(i) < L2(j), L1(j) > L2(i)\}|}{\left(\frac{1}{2}\right)|l|(|l|-1)}$$

It is the count of number of pairwise disagreements between two lists.

2.2.2 Spearman footrule distance

Given two full lists L1 and L2, each of cardinality |L|, the Spearman footrule distance is given by

$$F(L1, L2) = \frac{\sum_{\forall i} L1(i) - L2(i)}{\left(\frac{1}{2}\right)(|l|^2)}$$

Example x: For L = {2,5,3,1,4}, L1 = {3,2,5,4,1} and L2 = {5,3,2,4,1} the Spearman footrule distance can be calculated as follows

$$F(L, L1) = \frac{|4-5| + |1-2| + |3-1| + |5-4| + |2-3|}{0.5 * 5^2} = \frac{6}{12} = 0.5$$

$$F(L, L2) = \frac{|4-5| + |1-3| + |3-2| + |5-4| + |2-1|}{0.5 * 5^2} = \frac{6}{12} = 0.5$$

$$F(L1, L2) = \frac{|5-5| + |2-3| + |1-2| + |4-4| + |3-1|}{0.5 * 5^2} = \frac{4}{12} = 0.33$$

2.2.3 Normalized aggregated distance

Given a set of k full lists as $\{L_1, L_2, ..., L_k\}$, the normalized aggregated Kendall distance of a list L with respect to the set of k full lists is given by

$$K(L, \{L_1, \dots, L_k\}) = \frac{\sum_{i=1}^{k} K(L, L_i)}{k}$$

and the normalized aggregated Spearman footrule distance of a list L with respect to the set of k full lists is given by

$$F(L, \{L_1, \dots, L_k\}) = \frac{\sum_{i=1}^{k} F(L, L_i)}{k}$$

Example X: Continuing from the previous example, the normalized aggregated Spearman footrule distance can be calculated as

$$F(L, \{L_1, L_2\}) = \frac{F(L, L_1) + F(L, L_2)}{k} = \frac{0.5 + 0.5}{2}$$

2.3 Rank Aggregation Problem

Given a set of lists $\{L_1, L_2, ..., L_k\}$, the rank aggregation is defined as the problem of generating a list L such that L is closest to $\{L_1, L_2, ..., L_k\}$.

2.3.1 Optimal Rank Aggregation

Given a set of lists $\{L_1, L_2,..., L_k\}$, the optimal rank aggregation is defined as the problem of generating a list L such that either $K(L, \{L_1, L_2,..., L_k\})$ or $F(L, \{L_1, L_2,..., L_k\})$ is minimized [1,2].

2.3.2 Partial footrule optimal aggregation(PFOA)

Given a set of partial lists $\{L_1, L_2,..., L_k\}$, the optimal rank aggregation is defined as the problem of generating a list L such that either $K(L, \{L_1, L_2,..., L_k\})$ or $F(L, \{L_1, L_2,..., L_k\})$ is minimized [2].

This is a special case of optimal rank aggregation where the lists are partial top d lists obtained from search engine results. We have retrieved the top 100 results from each search engine and then calculated the partial footrule optimal aggregation using various algorithms in the following chapters.

Chapter 3

Rank Aggregation Algorithms

3.1 Positional Methods

The methods based on the position of elements are called positional methods. Following is the most popular positional method found in literature.

3.1.1 Borda's Method for Rank Aggregation

A popular and old rank aggregation method is Borda's method [1,2,5]. It calculates a score of each element using the positions of the elements in the ranked lists. It is a common technique used in voting scheme, where each candidate receives a score by each voter, the candidate with the highest score attains the first rank, the second highest attains the second rank and so on. In the context of web meta-searching each document of a list corresponds a candidate and each voter corresponds to a search engine result. Thus, each search engine rates each document according to its algorithm and produces a result in the form of a lists of documents. For example, having N search engines results having d element each, a document receives a score d-1 for a search engine if it tops all the d documents of the search engine, it gains a cumulative score of (d-1)*N if it is at top position in all the search engines. Similarly, it receives a score of d-2 for a search engine if it has rank 2 in that search engine result and so on. Combining all the scores from each N search engines for a document a final score is obtained which is called as Borda Score. A similar final score is obtained for each of the d documents. Finally, we have a list of score of all the documents. Taking the descending sort on the Borda Score of all the documents the aggregated list is obtained.

Borda's method can be expressed mathematically as follows,

Given a set of k lists $\{L_1, L_2, ..., L_k\}$, to each element c_j in L_i, the Borda's Method assigns a score given as

$$B_i(c_j) = |c_p: L_i(c_p) > L_i(c_j)|$$

The total score of each element is given by

$$B(c_j) \sum_{i=1}^{k} B_i(c_j)$$

The descending sort on the Borda's Score gives the aggregated list [2].

3.2 Heuristic based methods

3.2.1 Mean by Variance

The two attributes of ranked documents namely, mean of positions and variance of positions are taken into consideration in order to calculate mean by variance. The mean of positions is the aggregate position of a document and variance is the statistical variance of the document based on positions and mean position.

It has been proved that the descending sort on the Borda Count is equivalent to ascending sort in mean of positions [2,5]. However, if the mean of positions of any two documents is found to be same, the document having larger variance should be given preference. Conversely, if the variance of positions of any two documents is found to be same, the document with smaller mean of positions should be preferred. In this way, combining the two principles a ratio mean by variance is utilized [2].

The ration mean by variance of a document di is calculated as

$$\frac{m}{v} = \frac{\overline{x}_{d_i}}{\overline{\sigma}^2_{d_i}}$$

A list *m/v(i)* for i = 1…M is obtained for M documents. The ascending sort on this list produces a corresponding sequence of positions of documents d_j for j Є [1,M]. This sequence of d_j's is the aggregated list of documents.

Example X: Given L1 = {3, 4, 2, 1}, L2 = {2, 4, 3, 1} and L3 = {4, 2, 1, 3}, the mean of positions of documents are $X_1 = (4 + 4 + 3)/3 = 3.67$, $X_2 = 2.0$, $X_3 = 2.67$, $X_4 = 1.67$, and the variance of positions are $\overline{\sigma}_1^2$ = [(4 − 3.67)2 + (4 − 3.67)2 + (3 − 3.67)2]/3 = 0.22, $\overline{\sigma}_2^2$ = 0.67, $\overline{\sigma}_3^2$ = 1.56, $\overline{\sigma}_4^2$ = 0.22. m/v(1) = (x1/ σ21) = 3.67/0.22 = 16.68, m/v(2) = 2.985, m/v(3) = 1.71, m/v(4) = 7.59. On applying the ascending order sort on m/v(i) for i = 1….M, we get m/v(3) < m/v(2) < m/v(4) < m/v(1). Hence, the aggregated list is L = {3, 2, 4, 1}.

This technique have been implemented and the results obtained from mean by variance have been displayed in Fig.1, Table 2 and Table 4 in Chapter 5.

While conducting the experiments it was observed that in case $\overline{\sigma}^2$ evaluates to zero, the technique fails due to division by zero. This failure can be observed in Table 1 and Table 4 for both the queries “zener” and “graphic design” in which all the search engines rank the document 1 to position 1 leading to variance of document 1 as zero. An improvement have been suggested in Chapter 4 whish shows what necessary measure should be taken so that the technique does not fail.

3.3 Fuzzy logic based methods

3.3.1 Membership Function Ordering

The application of mean and variance of position in mean by variance technique yields fruitful results [2]. While applying the same two attributes in fuzzy logic, the Gaussian membership function can be obtained as

$$\mu_{\tilde{d}_i}(x) = \frac{1}{\sqrt{2\pi\ \sigma_{d_i}^2}} \exp\left(-\frac{1}{2}\left[\frac{(x - \overline{x}_{d_i})^2}{\sigma_{d_i}^2}\right]\right)$$

Where $\overline{x}_{d_i}$ is mean of documents and $\sigma_{d_i}^2$ is variance of documents.

3.3.2 Shimura's Technique

Shimura had introduced a fuzzy algorithm for rank ordering of objects [14]. Let U be the universal set of objects.

The pairwise function $f_y(x)$ is a membership function of object x with respect to y, with $f_x(x) = 1$ and $f_y(x) = 0$ for $x \neq U$.

The relativity function f(x|y) is the fuzzy measurement of choosing x over y is defined as

$$f(x|y) = f_y(x) / \max[f_y(x), f_x(y)]$$

Now, for a given set $X=(x_1, x_2, x_3,\ldots,x_n)$, and the object $x \in X$ the relativity function, for choosing x among $X=(x_1, x_2, x3,\ldots.,x_n)$ is defined as

$$f(x|X) = \min\{ f(x|x_1),\ldots,f(x|x_n) \},\ f_x(x) = 1$$

Shimura's technique applied in meta-searching [2] is as follows,

In, meta-searching we have results of N search engines, resulting into N lists, $L_1,\ldots.L_N$, such that each list is a subset of U.

For k = 1,…,N and i, j = 1,…,n, the pairwise function can be defined as

$$f_{x_j}(x_i) = \frac{|L_k(x_i) < L_k(x_j)|}{N}$$

where $L_k(x)$ is the rank of element x in list L_k.

For n elements each in N lists, the relativity function of each element with respect to each other element, can be defined as

$$C_i = \min_{j=1..n}\{f(x_i|x_j)\}$$

$$= \min_{j=1..n}\{ fx_j(x_i) / \max[fx_j(x_i), fx_i(x_j)] \}$$

Now, C_i is the membership value of rank of each element. The descending sort on C_i gives the aggregated list.

3.3.3 Modified Shimura's Technique

The membership value of rank of each document is determined using Shimura's algorithm which employs min() function for choosing x among n elements. Beg observed that due to min() function the descending sort on Ci results in many ties. Therefore, the min() function was replaced by an OWA operator. To calculate the weights of OWA operator, the "at least half" linguistic quantifier was applied as follows,

For a relative quantifier with m criteria, the weights can be calculated as,

$$w_i = Q\left(\frac{i}{m}\right) - Q\left(\frac{i-1}{m}\right), i = 1,2,3,\dots,m \; with \; Q(0) = 0$$

where Q(r) is the membership function of the relative quantifier, defined as,

$$Q(r) = \begin{cases} 0 \; if \; r < a \\ \dfrac{r-a}{b-a}, if \; b \; \leq r \; \leq a, \\ 1 \; if \; r > b \end{cases}$$

For, at least half quantifier, a = 0, b = 0.5.

The improved Shimura is yields better results than Shimura.

Chapter 4

Improved Rank Aggregation Algorithms

4.1 Variance by mean (Improved Mean By Variance)

As stated earlier, in mean by variance technique, the problem of division by zero arises when $\overline{\sigma}^2$ evaluates to zero and no solution have found in literature to deal with this problem. Therefore, we propose an improvement which is simple and robust to the division by zero error.

Instead, of calculating the ratio of mean of positions to variance of positions, we calculate its inverse, that is, the ratio of variance of positions to mean of position.

The ration mean by variance of a document di is calculated as

$$\frac{v}{m} = \frac{\overline{\sigma}^2{}_{d_i}}{\overline{x}_{d_i}}$$

A list *v/m(i)* for i = 1…M is obtained for M documents. The descending sort on this list produces a corresponding sequence of positions of documents d_j for j Є [1,M]. This sequence of d_j's is the aggregated list of documents.

Example X: Given L1 = {3, 4, 2, 1}, L2 = {2, 4, 3, 1} and L3 = {4, 2, 1, 3}, the mean of positions of documents are $X_1 = (4 + 4 + 3)/3 = 3.67$, $X_2 = 2.0$, $X_3 = 2.67$, $X_4 = 1.67$, and the

variance of positions are $\overline{\sigma}_1^2$ = [(4 − 3.67)2 + (4 − 3.67)2 + (3 − 3.67)2]/3 = 0.22, $\overline{\sigma}_2^2$ = 0.67, $\overline{\sigma}_3^2$ = 1.56, $\overline{\sigma}_4^2$ = 0.22. v/m(1) = ($\overline{\sigma}_1^2$/x1) = 0.22/3.67 = 0.0599, v/m(2) = 0.335, v/m(3) = 0.5842, v/m(4) = 0.1317. On applying the ascending order sort on m/v(i) for i = 1….M, we get v/m(3) > v/m(2) > v/m(4) > v/m(1). Hence, the aggregated list is L = {3, 2, 4, 1}.

The results of variance by mean have been displayed comparatively along with the results of the other techniques implemented and it is revealed that it works as good as Shimura's technique.

4.2 Enhanced Shimura

In modified Shimura technique, it has been observed that the membership function of the fuzzy relative linguistic quantifier "at least half" is piecewise linear in nature, due to which, while calculating the weights from membership values half of the m weights become zero, that is,

$$Q(r) = 1 \text{ for all } r \geq 0.5,$$

Hence,

$$w_i = Q\left(\frac{i}{m}\right) - Q\left(\frac{i-1}{m}\right)$$

$$= 1 - 1,$$

$$= 0, \quad \text{for } i = \lfloor m/2 \rfloor, \lfloor m/2 \rfloor + 1, \ldots, m$$

Therefore, in order to satisfy the Condorcet criteria, ignoring the membership values of lower half elements by taking zero weights is not pertinent.

The solution could be the choice of a an appropriate membership function which gives the values of weights in decreasing fashion as moving from membership value 0 to 1. This also satisfies the Condorcet criteria. It can only be achieved by a non-linear membership function.

Therefore, the conclusion is that the results depends upon the choice of appropriate membership function used for the calculation of weights of OWA operator.

By intuition, in order to achieve similar results, the membership function that is similar to the piecewise linear "at least half" relative quantifier is √x. The other choice of a non-linear function may be x^2.

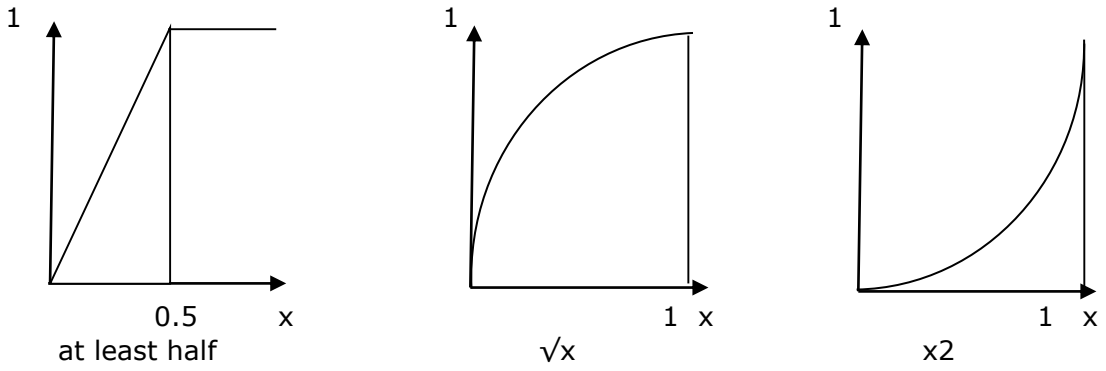


Fig. 4.1 Linguistic quantifier "at leat half" and proposed fuzzy quantifiers

The desired membership function of the quantifier can be defined as,

$$Q(r) = \begin{cases} \sqrt{r} & \text{if } 0 \leq r \leq 1, \\ 0 & \text{otherwise,} \end{cases}$$

Example: For m = 5, we have Q(0) = 0, Q(1/5) = 0.4472, Q(2/5) = 0.6325, Q(3/5) = 0.7746, Q(4/5) = 0.8944, Q(5/5) = 1, therefore we get, w1 = Q(1/5) - Q(0) = 0.4472, w2 = 0.1853, w3 = 0.1421, w4 = 0.1198, w5 = 0.1056.

It can be observed from the previous example that the weights obtained in the above example are in desired values.

Better results than the improved Shimura are experimentally achieved by using the non-linear membership function √x and x2 in place of linguistic quantifiers when applied on full lists. They have to be tested on real data of partial lists.

Chapter 5

Experiments and Results

5.1 Experimental setup

We have collected the search engine results of Bing, Gigablast, Google, MySearch and Wow and developed a benchmark dataset. The queries executed on these search engines are the subset of the queries used by [1,2] are as follows:

Affirmative action, alcoholism, amusement parks, architecture, bicycling, blues, cheese, citrus groves, classical guitar, computer vision, cruises, Death Valley, gardening, graphic design, gulf war, HIV, lyme disease, mutual funds, parallel architecture, sushi and zener.

We have collected top-100 results from each search engine returned by each query. The average number of search results per query is 327. The minimum and maximum being 293 and 369 respectively. We have omitted the video links and only considered the links of text documents viz. html, pdf etc. The google rankings are taken as base ranking starting from 1 to 100, and then adding one to each new result from any of the search engine.

We have examined all the algorithms discussed in the previous chapters.

5.2 Results

5.2.1 Results of Variance by Mean(VBM)

The Table 1 shows the top 5 results of query zener in the form of document identifiers from all the five search engines as well as the aggregation results from VBM and Borda's method. It is observed that all the search engine rank the document 1 at position 1. With this observation,

the value of mean of document 1 evaluates to 1 and the value of variance for the same evaluates to 0, leading to the failure of mean by variance technique. This drawback is addressed and removed in variance by mean technique. The Normalized Aggregated Footrule Distance of VBM and Borda are 0.4863 and 0.4347 respectively. Although the VBM performs poor than Borda's method, the technique is robust as compared to mean by variance which fails when variance of any document is zero.

Search Engine Results					Aggregation Results	
Bing	Gigablast	Google	MySearch	Wow	VBM	Borda
1	1	1	1	1	2	1
112	226	2	333	2	4	5
113	227	3	334	4	10	13
114	228	4	5	5	14	7
78	229	5	2	6	22	23

Table 5.1: Top five results of query Zener

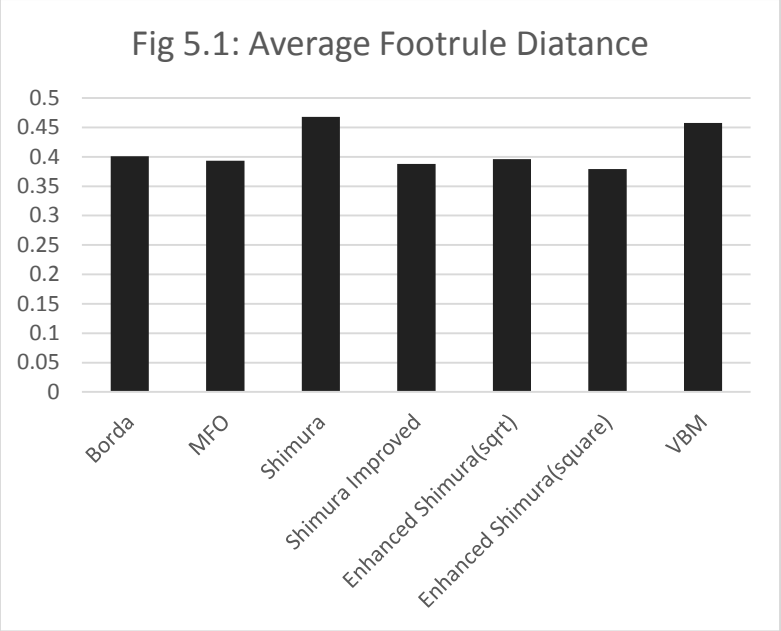
5.2.2 Results of Enhanced Shimura Technique

The Table 2 represents the Footrule Distances obtained from all the 21 queries mentioned previously. It can be observed that the Enhanced Shimura Technique using square membership function shows the best performance over all the techniques almost in all the queries.

The average Normalized Footrule Distance has been represented in the Fig. 2. It clearly shows that the Enhanced Shimura Technique using square membership function outperforms all the techniques. Improved Shimura bags the second place followed by Membership Function Ordering.

The Enhanced Shimura Technique using the sqrt membership function performs better than Borda's Method.

The Table 3. shows the top 10 results of all the search engines as well as those of fuzzy logic aggregation techniques. The results of Enhanced Shimura(square) are all most same as those of Improved Shimura and Enhanced Shimura(sqrt), but with a difference at position 7 and position 8 yielding to better performance.



S.No.	Query	|U|	Normalized Footrule Distance						
			Borda	MFO	Shimura	Shimura Improved	Enhanced Shimura Improved (sqrt)	Enhanced Shimura Improved (square)	VBM
1	affirmative action	313	0.4084	0.3916	0.4765	0.3912	0.3962	0.3796	0.4687
2	alcoholism	336	0.3805	0.3719	0.4669	0.3806	0.379	0.3814	0.4558
3	amusement parks	357	0.4023	0.3796	0.4354	0.3927	0.3947	0.386	0.4577
4	architecture	344	0.4079	0.3847	0.4481	0.3871	0.3977	0.375	0.4775
5	Bicycling	315	0.4127	0.4219	0.5044	0.3978	0.4015	0.3903	0.4341
6	Blues	338	0.4131	0.407	0.4611	0.3992	0.4117	0.389	0.4482
7	Cheese	369	0.3945	0.3889	0.452	0.3909	0.3954	0.381	0.4328
8	citrus groves	318	0.3955	0.401	0.4931	0.3771	0.3850	0.372	0.4387
9	classical guitar	319	0.3898	0.3845	0.4778	0.3806	0.3870	0.3725	0.4583
10	computer vision	293	0.3954	0.3999	0.5019	0.3832	0.3952	0.3643	0.4848
11	Cruises	307	0.4146	0.4195	0.4556	0.4043	0.4138	0.3957	0.4571
12	Death Valley	299	0.3818	0.3694	0.4463	0.3632	0.3712	0.3524	0.479
13	gardening	331	0.3878	0.375	0.4759	0.3738	0.3843	0.3598	0.4423
14	graphic design	357	0.4041	0.3871	0.4787	0.3923	0.4013	0.3847	0.4425
15	gulf war	311	0.4057	0.3869	0.4827	0.3847	0.3924	0.3789	0.4765
16	HIV	319	0.4031	0.4008	0.4957	0.3896	0.3936	0.385	0.4496
17	lyme disease	314	0.3918	0.3869	0.4489	0.3756	0.3893	0.3656	0.4532
18	mutual funds	350	0.3989	0.4067	0.4417	0.4016	0.4043	0.3926	0.4802
19	parallel architecture	322	0.3953	0.393	0.461	0.3817	0.389	0.3773	0.4333
20	Sushi	338	0.4014	0.3937	0.4399	0.3858	0.3968	0.3774	0.4516
21	Zener	320	0.4347	0.4043	0.4814	0.4132	0.4323	0.3987	0.4863

Table 5.2: Query wise Performance of Rank Aggregation Techniques based on Normalized Footrule Distance

Aggregation Technique	Time relative to Borda
Borda	1
MFO	537.0248
Shimura	13.15276
Shimura Improved	4.206506
Enhanced Shimura(sqrt)	3.011315
Enhanced Shimura(square)	2.956860
VBM	3.968317

Table 5.3: Comparison of time taken in relative to Borda

Search Engine Results					Aggregation Results of Fuzzy Logic Techniques			
Bing	Gigablast	Google	MySearch	Wow	MFO	Improved Shimura	Enhanced Shimura (sqrt)	Enhanced Shimura (square)
1	1	1	1	1	1	1	1	1
5	247	2	354	3	10	3	3	3
8	249	3	5	150	28	5	5	5
112	250	4	35	445	8	8	8	8
55	251	5	150	6	7	150	150	150
113	252	6	455	7	150	10	10	10
33	253	7	58	8	26	7	7	6
114	254	8	355	9	48	6	6	7
116	255	9	53	10	33	9	9	9
117	256	10	15	11	55	28	28	28

Table 5.4: Top 10 results of query graphic design

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this dissertation we have implemented new meta-searching algorithms successfully in order to provide web users useful and reliable results while searching the web. The motivation being that no search engine till date is sufficient and reliable to achieve the user satisfaction. We discussed the application of rank aggregation in meta-searching and also highlighted the issues and challenges faced to implement meta-searching using rank aggregation. Borda's method of rank aggregation has been implemented and considered as a benchmark for evaluation of proposed algorithms. We have successfully implemented and evaluated the proposed techniques which are improvements in the previous techniques. We have improved the mean by variance technique and developed variance by mean technique, improved the modified Shimura and developed two versions of enhanced Shimura. We have also implemented the old techniques namely mean by variance, membership function ordering, Shimura's technique for fuzzy ordering and modified Shimura technique and put forward a comparative analysis of old and proposed techniques. The proposed technique of Enhanced Shimura have been found most performance effective and time efficient against the other techniques significantly.

6.2 Scope for future work

The spam fighting can be tested for the proposed techniques in future work in order to make them more efficient. It can be performed by testing the techniques on the Condorcet criteria and then evaluating the performance.

References

[1] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the Web". In Proceedings of the Tenth ACM International Conference on World Wide Web, pages 613-622, 2001.

[2] M. M. Sufyan Beg and N. Ahmad, "Soft Computing Techniques for Rank Aggregation on the World Wide Web,". In World Wide Web Journal: Internet and Information Systems, volume 6, pages 5-22, 2003.

[3] M. M. Sufyan Beg and N. Ahmad, "Fuzzy Logic and Rank Aggregation for the World Wide Web". In Studies in Fuzziness and Soft Computing Journal, volume 137, pages 24-46, 2004.

[4] S. Yasutake, K. Hatano, E. Takimoto, M. Takeda "Online Rank Aggregation". In Proceedings of 24th International Conference ALT, pages 68-82, 2013.

[5] M. E. Renda and U. Straccia, Web Metasearch: Rank vs. Score Based Rank Aggregation methods. In Proceedings of the ACM Symposium on Applied Computing, March 09-12, 2003.

[6] L. Akritidis, D. Katsaros and P. Bozanis, "Effective Rank Aggregation for Meta searching". In The Journal of Systems and Software, volume 84, pages 130-143, 2010.

[7] T. Qin, X. Geng and T. Y. Liu, "A New Probabilistic Model for Rank Aggregation". In Proceedings of Advances in Neural Information Processing Systems 23, pages 681-689, 2010.

[8] J. A. Aslam and M. Montague, "Models of Metasearch". In Proceedings of 24th SIGIR, pages 276-284, 2001.

[9] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, Rank aggregation Revisited. Manuscript 2001.

[10] J. C. Borda, "Memoire sur les election au scrutin," Histoire de l'Academie Royale des Sciences, 1781.

[11] Y. T. Liu, T. Y. Liu, T. Qin, Z. M. Ma and H. Li, "Supervised Rank Aggregation". In Proceedings of the ACM International Conference on World Wide Web, pages 481-489, 2007.

[13] N. Ailon, "Aggregation of partial rankings, p-ratings and top-m lists". In Algorithmica, volume 57(2), pages 284-300, 2008.

[14] M. Shimura, "Fuzzy Sets Concepts in Rank Ordering Objects". In Journal of Mathematical Analysis and Applications, volume 43, pages 717-733, 1973.