ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 4, Special Issue 9, July 2015

National Conference on Emerging Technology and Applied Sciences-2015 (NCETAS 2015)

On 21st & 22nd February, Organized by

Modern Institute of Engineering and Technology, Bandel, Hooghly-712123, West Bengal, India

Meta Search Engine using Multi-Objective Partial Rank Aggregation: Application in Ranking WebPages

Sheuli Maity

Assistant Professor, Department of CSE, Modern Institute of Engineering & Technology, Bandel, West Bengal, India

ABSTRACT*:* Although there are hundreds of search engines no single search engine can satisfy all web users and can be considered broadly acceptable that Sufficiently comprehensive in its coverage of the web moreover they consist the "spam pages" when a web page gets an undeservedly high rank. Therefore, a robust technique for Meta Search Engine is required that can effectively combat "spam pages", a serious problem in Web searches. When some top ranked WebPages are extracted from different Search Engines, they differ each other, and the problem become more difficult to aggregate them. In this article a multi objective genetic algorithm based rank aggregation method has been proposed where some partial rankings are integrated in an unbiased way. The total distance from the reference ranking to the input rankings is minimized as the first objective. For distance calculation the scaled footrule distance is used. The standard deviation between those distances is minimized as the second objective in order to avoid biasness toward a particular input ranking. The proposed method has been applied on some most-usable Search Engines. Thereafter, the resultant ranking is compared with the other rank aggregation methods to check how much it is applicable to reduce the spam pages.

KEYWORDS: Partial list, Scaled Footrule Distance, Pareto optimal Front, Rank aggregation, Spam Page Reduction.

I.INTRODUCTION

Since today there are hundreds of general purpose search engines as well as special purpose search engines. But The fact that so many choices cannot satisfy all web users yet and hence can be considered broadly acceptable that sufficiently inclusive in its coverage of the web. The very first question arises for the spam pages [16] when a web page gets an undeservedly high rank Sometimes it is seen that some webpages are not workable but occupy a good rank in some search engines [3] [6]. Same webpages can be noticed to occupy more than 3 spaces in the first 50 ranks. Thus our first motivation is to find robust techniques for meta-search to provide users a certain degree of robustness of search to clarify various shortcomings and biases.

Secondly users must have the protection against the biases of individual Search engines [1] because the world is governed by commercial interests. The first reason is more challenging in doing meta-search as all the search engines will be capable of ranking the entire collection of pages on the Web, which is growing day to day. Secondly, Search Engines routinely limit access to about the first few hundreds of pages in their rank-ordering to ensure the confidentiality of their ranking algorithm and the efficiency. The issue of efficiency is also a serious bottleneck in performing rank aggregation for multi-criteria selection and word association queries [16].

All the WebPages cannot be covered by individual Search Engine at the same time. To compensate the shortage of these ranking schemes, all the rankings [7], [8], [9], [10] must be aggregated. Previously, different others rank aggregation methods like robust rank aggregation [11], voting rank aggregation [5], MCT [12], MC4 [12], [13], Borda [4], [13] etc have been studied. In case of full list, where all the rank lists contain all the WebPages, the task of aggregation is relatively easier. But, integrating partial lists which contain unequal number of WebPages is really a challenging issue. In this article, a multi-objective genetic algorithm [14], [15] based rank aggregation method for partial lists of ranks is proposed. Since this time, only minimum distance from the input rankings is considered in case of rank aggregation method. But our motivation is to remove the biasing property, hence consider also the standard deviation between the input rankings. Instead of single objective optimization which yields a single best solution, a multi-objective optimization

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 4, Special Issue 9, July 2015

National Conference on Emerging Technology and Applied Sciences-2015 (NCETAS 2015)

On 21st & 22nd February, Organized by

Modern Institute of Engineering and Technology, Bandel, Hooghly-712123, West Bengal, India

[16] produces a set of solutions that contains a number of Pareto-optimal solutions. Moreover multi-objective optimization problem has been modelled by GA in which fitness comparison takes Pareto dominance into account and non-dominated solutions are stored in an archive to approximate the Pareto front. The first objective is to minimize the total distance of the reference ranking (encoded in a chromosome of GA) from the input rankings. The other objective is to minimize standard deviation between those distances to avoid the biasness to a particular input ranking. For calculating the distances between two rankings Scaled Footrule distance measure [13] has been used. The proposed method has been applied on some most-usable Search Engines. Thereafter, the resultant ranking is compared with the other rank aggregation methods to check how much it is applicable to reduce the spam pages.

II.MULTI-OBJECTIVE OPTIMIZATION

When multiple goals or objectives are optimized simultaneously then the problem is referred as multi-objective optimization (MOO). MOO may estimate different aspects of solutions which are partially or wholly in conflict. Find the vector $\vec{x}^* = (x_1^*, x_2^*, \dots, x_n^*)^T$ of decision variables which satisfies m inequality constraints:

$$g_i(\vec{x}) \geq 0, \quad i = 1,2,\dots,m \tag{1}$$

P equality constraints

$$h_i(\vec{x}) = 0, \quad i = 1,2,\dots,p \tag{2}$$

And optimizes the vector function

$$\vec{\mathcal{F}}(\vec{x}) = [f_1(\vec{x}), f_2(\vec{x}), \dots, f_k(\vec{x})]^T \tag{3}$$

Here, $\mathcal{F}$ denotes the feasible region of the problem (i.e., where the constraints are satisfied). Any solution outside this region is unacceptable though it oversteps one or more constraints. The vector $\vec{x}^*$ denotes an optimal solution in $\mathcal{F}$. A Multi-objective searching concept is clearly described in [17]. More than one solution can be obtained in Pareto optimal set as there exist different 'trade-off' solutions to the problem. The set of solutions contained by Pareto optimal set are called non-dominated solutions and they create the Pareto front. Specifically MOO is a process of generating the whole Pareto front or an approximation to it.

III.CALCULATING THE DISTANCE

Generally, all ranking schemes are aggregated to make a consensus among them [2]. Aggregating a partial list of ranks having unequal numbers of features, i.e. where all the features are not present, is more challenging than full list. Induced Footrule Distance and Scaled Footrule Distance [13], [18] are mostly used in case of partial lists. Here we have used Scaled Footrule Distance (SFD). If σ is a full list and τ_i is partial list then the distance $SFD(\sigma, \tau_1, \tau_2, \dots, \tau_k)$ can be calculated as in Eq. 4. For normalizing the distance we just divide the SFD by $|\sigma^2|$.

$$SFD(\sigma, \tau_1, \tau_2, \dots, \tau_k) = \frac{\Sigma_{i\epsilon\tau}\left|\frac{\sigma_i}{|\sigma|} - \frac{\tau_i}{|\tau|}\right|}{\left|\frac{I}{2}\right|} \tag{4}$$

IV.GENERATING INPUT RANKINGS

Input rankings are generated by different Search Engines such as google(GG), yahoo(YH), hotbot(HB), bing(BG), lycos(LY) and some specific word like 'cheese', 'gardening', 'ship', 'education' etc. have been searched against them. Then the first 50 URL are collected against each word in respect of each Search Engine. This list of ranks is considered as the input file for aggregation. For each search, the WebPage differs, as same WebPage cannot be present in first 50 ranks of different Search Engines. Now the goal is to get the first 50 WebPages after aggregating all input lists. Same WebPages occupy more than 3 spaces in the first 50 ranks. First, the unique WebPages are extracted from the input lists and converted into full list. Here, an example of converting two partial lists to two full lists is shown in Figure 1. Two partial lists Partial list-1 and list-2 are generated with different ranking schemes and it can be seen that different WebPages are present in these two lists. Then, a unique collection of WebPages from these two lists are taken and the total number of unique WebPages is 14. Now, each list is converted into full list by giving all the unique WebPages a rank. If a WebPage is already present in existing partial list, then that WebPage is rank according to that ranking scheme. Such as, W1 is already present in list-1 as rank 7, therefore, in full list-1, the rank of W1 is 7. If WebPages which are not present in existing partial list, then that WebPages are ranked same and other than existing rankings. Such as, W5 is not present in list-1, so, in full list-1, W5 is ranked 15. Thus, two partial lists are transformed into two full lists and used as input data file for aggregation.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 4, Special Issue 9, July 2015

National Conference on Emerging Technology and Applied Sciences-2015 (NCETAS 2015)

On 21st & 22nd February, Organized by

Modern Institute of Engineering and Technology, Bandel, Hooghly-712123, West Bengal, India

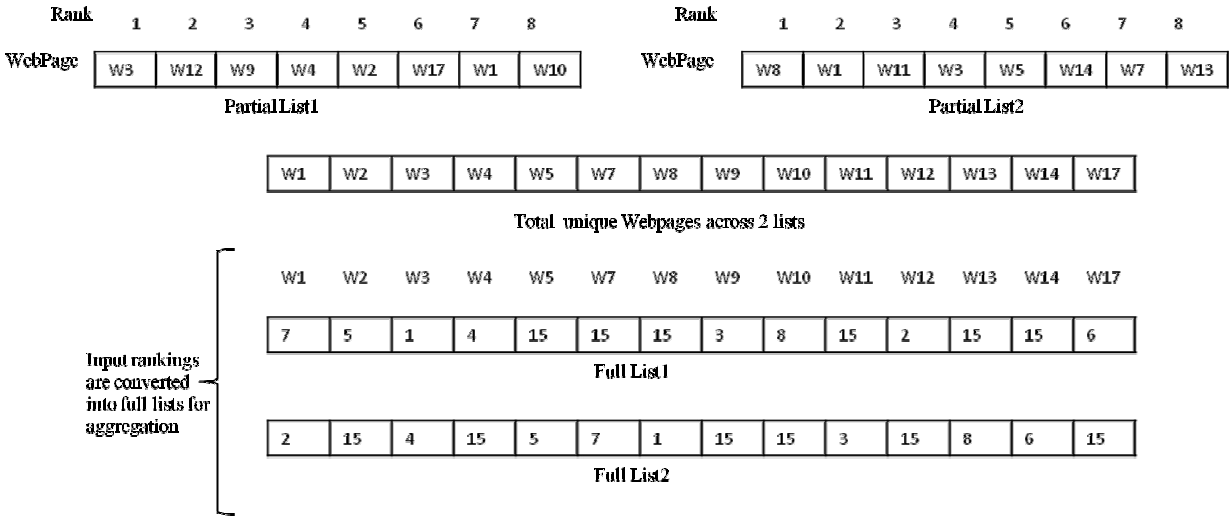

FIG. 1. : PARTIAL INPUT RANKINGS ARE CONVERTED INTO FULL LISTS

V.MULTI-OBJECTIVE GENETIC ALGORITHM (MO-GA) BASED RANK AGGREGATION

MO-GA is a procedure of multiple criteria optimal decision making in the presence of trade-offs between different conflicting objectives. Sometimes, a single solution is not enough to optimize each objective functions simultaneously, rather, a set of non-dominated solution i.e. pareto optimal front must be generated. Here two fitness functions are used to get an unbiased aggregated ranking. In each generation, total distance from reference ranking to input rankings along with the standard deviation between them are minimized.

1. ENCODING SCHEME & POPULATION INITIALIZATION

A population in genetic algorithm consists of P chromosome and each chromosome represents a reference ranking. If there are N unique WebPages in input ranking, then the size of a chromosome is N. As ranks must be integer, so integer encoded chromosome are used. Figure 2 illustrates the encoding scheme for the above input rankings described in section IV. As in the example 14 unique WebPages are present, so the length of the chromosome is 14. Firstly, randomly generated integer numbers encodes a chromosome, and gradually it improves with respect to fitness in each iteration.

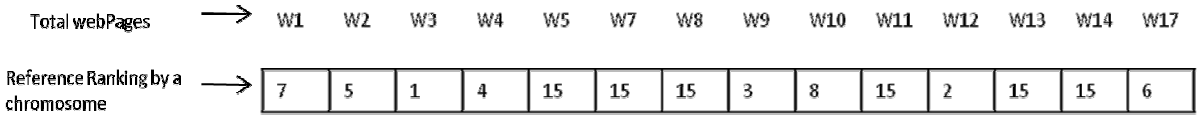

Fig. 2 Encoding Scheme

2. FITNESS CALCULATION

Accuracy depends on unbiased selection. That can be assured by considering the standard deviation among the distances. Two fitness functions have been considered here for multi objective optimization. Along with minimizing the total distance, standard deviation has also been minimized. Scaled Footrule distance measure has been chosen here for distance calculation. First, the distances from the reference ranking to the input rankings have been calculated using Eq. 4 and then mean distance is computed in Eq. 5. Here N is number of input rankings and R is the reference ranking. SFD is the Scaled Footrule Distance between two rankings. Secondly, the Standard Deviation between these distances has been calculated in Eq. 6.

$$Dist = \Sigma_{i=1}^{N}(R, Input - rank - order(i)) \quad (5)$$

$$SD = std(SFD_1, SFD_2, \dots, SFD_N) \quad (6)$$

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 4, Special Issue 9, July 2015

National Conference on Emerging Technology and Applied Sciences-2015 (NCETAS 2015)

On 21st & 22nd February, Organized by

Modern Institute of Engineering and Technology, Bandel, Hooghly-712123, West Bengal, India

3. SELECTION, CROSSOVER & MUTATION

Binary tournament selection [17] is chosen for selection procedure. Figure 3 describes an example of single-point crossover where cross point is 3. Firstly, the positions are calculated for each chromosome, as they represent the rank orders, but the crossover should be done against the WebPages. Secondly, from random selection, if a number is greater than the crossover probability, can be chosen for cross point. From parent1 sequence, the first fragment i.e. before the crossover point is directly copied to child1 i.e., [W2,W4,W1]. Then the set difference from parent1 [W2,W4,W1,W5,W3,W6] and first fragment of child1 [W2,W4,W1] is calculated as [W5,W3,W6] and let this set is denoted as S. Then the order of the set S are checked against in parent2 [W3,W1,W6,W4,W2,W5] for order. Now the set S in this order [W3,W6,W5] is put into the next part i.e. after crossover point of the child1. For the child2 this steps are repeated in vice versa. Again, the positions are calculated for each chromosome for the rank orders. For the mutation, 2 random numbers are selected and swapping has been done between the values of those indices (Fig. 4). Crossover probability, Mutation Probability is chosen as 0.9 and 0.1 respectively here.

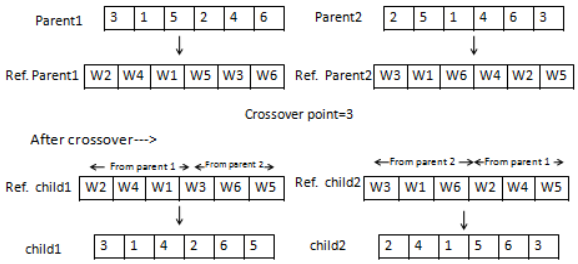

Fig. 3. : Crossover procedure

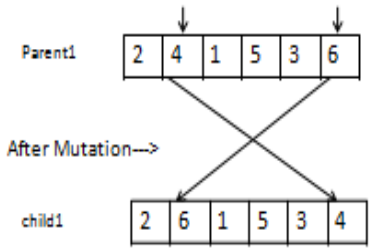

Fig. 4. : Mutation procedure

VI.RESULTS AND DISCUSSIONS

Here first 50 ranked pages are collected from google(GG), yahoo(YH), hotbot(HB), bing(BG), lycos(LY) against some specific word search like ‘cheese’, ‘gardening’, ‘education’. An appropriate rank is employed to them and converted into full list. The proposed method is compared with that of Voting (V), Robust ranking (R), Borda method (B) and two Markov Chain based approach MCT, MC4. They are compared to check how they reduce the spam pages in Search Engines.

For each of these queries, we give more attention at the (top) pages that we consider spam. The definition of “spam” does not mean malicious! --- these pages just obtained an unjustifiably high rank from other Search Engines. Sometimes it is seen that some WebPages are not workable but occupy a good rank in some Search Engines. Also, same WebPages occupy more than 3 spaces in the first 50 ranks. As a result, some well-known WebPages get very poor rank in some Search Engines whereas the same get very high rank in other Search Engines. If a WebPage is present less than the half of the total Search Engines, they are called ‘spam pages’. It is easy to find URLs that spammed a single Search Engine. When the Search Engines was not highly overlapped, the most challenging objective is to find those URLs that spammed at least two Search Engines. Table I presents our examples: the entries are the rank within individual Search Engines' lists. A blank entry in the table indicates that the URL was not returned as one of the top 50 by the Search Engine. From the table, it is clear that the pages ranked only one or two Search Engines out of 5, gets very poor rank in MOGA, considering them not to be spammed. But, the pages that are well ranked in at least 3 out of 5 Search Engines get average ranking in MOGA, better than the other existing method. Hence it is reducing spam.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 4, Special Issue 9, July 2015

National Conference on Emerging Technology and Applied Sciences-2015 (NCETAS 2015)

On 21st & 22nd February, Organized by

Modern Institute of Engineering and Technology, Bandel, Hooghly-712123, West Bengal, India

TABLE I.
COMPARISON BETWEEN THE RANKINGS OF SOME WEBPAGES AMONG THE PROPOSED AND EXISTING METHODS

WebPage	GG	Y H	HB	BG	LY	V	R	MCT	MC4	B	MOGA
http://www.x2d.org/misc/misc3/wallsFixed.html	15					65	50	50	48	1	147
https://live.gnome.org/Cheese		24	19	20	19	7	62	164	164	13	14
http://www.stiltoncheese.co.uk/				16	17	23	54	53	52	5	142
http://www.ilovecheese.com/	5			9	11	6	23	22	20	142	15
http://www.bigcheesecoaching.com/	33	34	32			30	99	164	164	50	65
http://www.gardening.com/	13	20		20		12	10	48	47	48	4
http://www.rd.com/home/gardening/				24		76	30	154	154	55	106
http://www.hgtvgardens.com/	26				27	49	104	154	154	88	26
http://www.quotegarden.com/gardens.html		47		50		137	50	154	154	135	126
http://www.bing.com/images/search?q=gardening &qpvt=gardening&FORM=IGRE		3		3	2	5	4	8	8	9	12
http://www.theguardian.com/lifeandstyle/gardens	22	28		31	23	15	12	154	154	80	69
http://www.burpee.com/	43					132	133	154	154	129	136

VII. CONCLUSION

In this article, rank aggregation method using multi-objective genetic algorithm has been projected to aggregate some partial rank lists of WebPages. The first intention is to minimize the total distance from the reference ranking to the input rankings and for distance calculation; the Scaled Footrule Distance is used. The second objective is to minimize the standard deviation between those distances to avoid biasness towards any particular input ranking. The proposed method has been applied on some most-usable Search Engines. Thereafter, the resultant ranking is compared with the other rank aggregation methods to check how much it is applicable to reduce the spam pages. Undoubtedly, the projected method performs well all over the Search Engines for reducing spam. We would like to extend our work in word association queries in web application in future.

REFERENCES

[1] D. E. Critchlow. "Metric Methods for Analyzing Partially Ranked Data", Lecture Notes in Statistics 34, Springer-Verlag, 1985.
[2] J. I. Marden. "Analyzing and Modeling Rank Data. Monographs on Statistics and Applied Probability", No 64, Chapman & Hall, 1995.
[3] M. Truchon. "An extension of the Condorcet criterion and Kemeny orders". cahier 98-15 du Centre de Recherche en Economie et Finance Appliquees, 1998.
[4] H. P. Young. "An axiomatization of Borda's rule". Journal of Economic Theory, 9:43--52, 1974.
[5] H. P. Young. "Condorcet's theory of Voting". American Political Science Review, 82:1231--1244, 1988.
[6] H. P. Young and A. Levenglick. "A consistent extension of Condorcet's election principle". SIAM J. on App.Mathematics, 35(2):285-300, 1978.
[7] R. Prati, "Combining feature ranking algorithms through rank aggregation," in 2012 International Joint Conference on Neural Networks,2012.
[8] Y.-T. Liu, T.-Y. Liu, T. Qin, Z.-M. Ma, and H. Li, "Supervised rank aggregation," International World Wide Web Conference Committee (IW3C2)., 2007.
[9] Q. Fang, J. Feng, and W. Ng, "Identifying differentially expressed genes via weighted rank aggregation," in IEEE International Conference on Data Mining, 2011.
[10] A. Veloso, M. A. Goncalves, and W. M. Jr., "Competence-conscious associative rank aggregation," Journal of Information and Data Management, vol. 2, no. 3, pp. 337–352, 2011.
[11] R. Kolde, S. Laur, P. Adler, and J. Vilo, "Robust rank aggregation for gene list integration and meta-analysis", Bioinformatics., 2012.
[12] R. P. DeConde, S. Hawley, S. Falcon, and et al., "Combining results of microarray experiments: A rank aggregation approach," Stat Appl Genet Mol Biol, vol. 5, no. 1, 2006.
[13] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the web," in Proceedings of the 10th international conference on World Wide Web, 2001.
[14] K. Deb, A. Pratap, S. Agrawal, and T. Meyarivan, "A fast and elitist multi-objective genetic algorithm: NSGA-II," IEEE Transactions on Evolutionary Computation, vol. 6, pp. 182–197, 2002.
[15] K. C. Mondal, A. Mukhopadhyay, U. Maulik, S. Bandhyapadhyay, and N. Pasquier, "Simultaneous clustering and gene ranking: A multi-objective genetic approach," in Proceedings of CIBB Palermo : Italy (2010), 2010.
[16] K. Deb, Multi-objective Optimization Using Evolutionary Algorithms. England: John Wiley and Sons, Ltd, 2001.
[17] C. A. C. Coello, "A comprehensive survey of evolutionary-based multi-objective optimization techniques," Knowledge and Information Systems, vol. 1, no. 3, pp. 129–156, 1999.
[18] P. Diaconis and R. Graham., "Spearmans footrule as a measure of disarray." Journal of the Royal Statistical Society, Series B, vol. 39, no. 2, pp. 262–268, 1977.