Computer Science and Engineering 2012, 2(5): 43-45
DOI: 10.5923/j.computer.20120205.01

Statistical Score Calculation of Information Retrieval Systems using Data Fusion Technique

B. Dhivakar[1], S. V. Saravanan[2], M. Sivaram[1], R. Abirama Krishnan[3]

[1]CSE Department, Anna University of Technology, Coimbatore, 64104, India
[2]Principal, Christ the King Engineering, College, Karamadai, Coimbatore, 641104,India
[3]Lecturer in the PG Department of Commerce with CA, HKRH College, Uthamapalayam, Theni, 625516, India

Abstract Effective information retrieval is defined as the number of relevant documents that are retrieved with respect to user query. In this paper, we present a novel data fusion in IR to enhance the performance of the retrieval system. The best data fusion technique that unite the retrieval results of numerous systems using various data fusion algorithms. The study show that our approach is more efficient than traditional approaches.

Keywords Rank Position, Borda Count, Condorcet Fusion, Fisher Criterion

1. Introduction

A retrieval system is a machine that receives the user query and generate the relevance score for the query-document pair. The process of finding the needy information from a repository is a non-trivial task[1-3] and it is necessary to formulate a process that effectively submits the pertinent documents. The process of retrieving germane articles[4] is termed as Information Retrieval (IR). It deals with the representation, storage, organization of and access to the information items[3]. Fusion is a technique that merges results retrieved by different systems to form a unique list of documents. Document Clustering is based on particular ranked list and does not take benefit of multiple ranked list. The fusion function accepts these score as its output for the query document pair. A static fusion function has only the relevance scores for a single query-document pair as its inputs. A dynamic fusion function can have more inputs. To construct a dynamic fusion function that can adjust the way it fuses multiple retrieval systems relevance scores for a query document pair using additional input features such as query, retrieved documents and joint distribution of retrieval systems relevance score for the query. Various models, schemes and systems have been proposed to represent and organize the document collection in order to reduce the users' effort towards finding relevant information[5]. In this study we present three different data fusion methods namely Rank Position, Borda Count, and Condorcet method in ranking retrieval systems. There are four feature selection techniques including Fisher Criterion, Golub Signal-to-Noise, traditional t-test and Mann-Whitney rank sum statistic.

2. Related Work

Fox and Shaw showed the five combination function for combining scores[6]. They are as follows:

CombMIN = Minimum of Individual Similarities
CombMAX = Maximum of Individual Similarities
CombSUM = Summation of Individual Similarities
CombANZ = CombSUM ÷ Number of non zero Similarities
CombMNZ = CombSUM × Number of non zero Similarities.

Fusion functions which are different from Comb- functions with respect to the generation of answer sets, are also found in the literatures[8]. These functions assign ranks to the documents in the answer set against the relevance score assignment mechanism adapted in Comb-functions. Few such fusion techniques which emulate the social voting schemes, are the Borda and Condorcet fusions[8]. Borda Fuse and Condorcet's fuse, and showed that the use of social welfare functions (Roberts, 1976) as the merging algorithms in data fusion generally outperforms the CombMNZ algorithm. Extensive work on Comb functions has been carried out by Lee[9–11] and based on the results he proposed few new rationales and indicators for data fusion. He concluded that CombMNZ is the better performing function than the others. The Probabilistic approach[12] differs from the Comb-functions in the way it selects a best performing strategy from a pool based on a predetermined probability value. The probabilistic model selects only one strategy from the pool while all other strategies remain unused. Hence, evolutionary algorithms are used to select the best performing strategies[13]. Meng and his co-workers (2002) indicate

* Corresponding author:
dhivabill@gmail.com (B. Dhivakar)

Published online at http://journal.sapub.org/computer

that metasearch software involves four components:

1. Database search engine selector: the search engines [database] to be mingle selected using some system selection methods
2. Query reporter: the queries are submitted to underlying search engines.
3. Document Selection tool: Documents to be used from each search engine are determined. The simplest way is the use of the top documents.
4. Unification of Result : The results of search engines are combined using merging techniques.

2.1. Lees Overlap Measure

Lee's [Lee, 1997] overlap measures, O_{rel} and O_{nonrel}, which measure the proportion of relevant and nonrelevant documents in the intersection of the two lists. These two measures are calculated as:

$$O_{rel} = \frac{2*I_{rel}}{R_1+R_2} \qquad (1)$$

$$O_{nonrel} = \frac{2*I_{nonrel}}{N_1+N_2} \qquad (2)$$

where R_i is the number of relevant documents and N_i is the number of nonrelevant documents returned by the system i respectively. The ratio of the two systems found to be an important predictive factor for the improvement of the combination. The similarity measure is the two systems on relevant document is less important than on relevant ones. After normalizing the scores for each system on each query by dividing their respective means we found the optimal combination for each possible. For each feature, we use one of the statistical methods such as the traditional t-test. Large score suggests that the corresponding feature has different expression levels in the relevant and irrelevant documents and thus is an important feature and will be selected for further analysis. Besides that some researchers used a variation of correlation coefficient to select features, for example Fisher Criterion [13] and Golub Signal-to-Noise.

3. Data Fusion Scheme for Determining Top Ranked Relevant Documents

3.1. Rank position method

The rank position of the retrieved documents are used to merge the documents into a single list . The rank position is determined by the retrieval system. We call d as the original document, while its counterparts in all other documents list are called Reference documents R_d of d. The following equation shows the statistical score calculation of document I using the position information of this document in all the systems (j=1,2,3,4…n).

$$r(d_j) = \frac{1}{\sum_j 1/position\ (d_{ij})} \qquad (3)$$

In this summation, systems not ranking a document are omitted. The unite of the top documents is treated as reproduced results.

3.2. Borda Count Method

Borda count and Condorcet method are based on democratic election strategies. The person with high score gets n votes and each successor gets one vote less than the predecessor i.e (n-1). If there are persons who are not interested in voting process, then the score is evenly divided among unranked candidates.Then, for each subsequent, all the votes are supplemented and the alternative with the highest number of score wins the election.

3.3. Condorcet method

In the Condorcet election method, voters rank the candidates in the order of partiality. It is a distinctive method that denote the winner as the candidate. Which prevail each of the other candidates in a pair-wise evaluation. To rank the documents we use their win and lose values.

4. Selection of Information Retrieval Systems for Data Fusion Technique

We consider three approaches for the selection of information retrieval systems to be used in data fusion.

Best: The best performing retrieval systems that achieve high percentage of the relevant documents retrieved are employed for statistical score calculation.

Normal : All systems to be ranked are used in data fusion.

Bias: The dissimilarity measure of the retrieval systems are used in data fusion.

The Fisher Criterion, Golub Signal-to-Noise , traditional t-test and Mann-Whitney rank sum statistic were applied to calculate the statistical score, S, for the IRs. In these techniques, each system was measured for correlation with the class according to some measuring criteria in the formulas. The systems were ranked according to the score, S, and the top ranked relevant documents in the IRs were selected. The Fisher Criterion, fisher is a measure that indicates how much the class distributions are separated. The coefficient has the following formula:

$$fisher = \frac{(\mu_1-\mu_2)^2}{(v_1+v_2)} \qquad (4)$$

Whereμ_i is the mean and v_i is the variance of the given IR whose documents are top ranked or otherwise in class i. There were two IR classes in this experiment, i.e. the relevant documents in IR and the non-relevant documents in IRs. The statistic gives higher scores to IR system that returns relevant document that are retrieved with respect to the user query, whose mean differ greatly between the two classes, relative to their variances.

Golub used a measure of correlation that emphasizes the "Signal-to-Noise" ratio, *signaltonoise*, to rank the relevant documents that are retrieved from the IRs. It is very similar to the Fisher Criterion but use another related coefficient formula as shown below:

$$signal\ to\ noise = \frac{(\mu_1 - \mu_2)^2}{(\sigma_1 + \sigma_2)} \tag{5}$$

Where μ_i is the mean and σ_i is the standard deviation of the relevant documents retrieved in class i.

Traditional t-test,ttest assumes that the values of the two class variances are equal. The formula is as follows:

$$ttest = \frac{(\mu_1 - \mu_2)}{\sqrt{((v_p/n_1)+(v_p/n_2))}} \tag{6}$$

Where μ_i is the mean of the relevant documents in class i and v_p is the pooled variance.

The Mann-Whitney rank sum statistics, mann has the following formula:

$$mann = \frac{n_1 * n_2 * (n_1+1)}{2 - r_1} \tag{7}$$

Where n_i is the sizes of class i, and r_1 is the sum of the ranks in class i. The score,S, for each relevant documents retrieved in the IR is thus calculated by using the formula in these statistical techniques.

The bias concept is used for the selection of IR system for data fusion. The cosine similarity measure is given by the following equation:

$$S_{AB}^{Cosine} = \frac{\sum_{i=1}^{N} A_i B_i}{\{\sum_{i=1}^{N} (A_i)^2 \sum_{i=1}^{N} (B_i)^2\}^{\frac{1}{2}}} \tag{8}$$

The bias between these two vectors is defined by subtracting the similarity value form 1.

$$B(A,B) = 1 - s(A,B) \tag{9}$$

We may use any of the combination of the above measure to calculate the statistical score of the information retrieval systems.

5. Discussion

So far, our study suggest that, for our choice of retrieval systems, there is an opportunity to improve the retrieval performance by fusing the above mentioned approach. Our preferred design of effective statistical score calculation of information retrieval systems is a multilayer technique to maximize precision and improve the retrieval performance that satisfies the user needs. In this paper we have summarized various methods that are used in different articles published in the journal thereby incorporating and integrating few of the approaches may lead to better precision and recall values. Our significant contribution is thereby invoking the methods thereby integrating few of the techniques from various research articles so that it will be useful to the researchers for their valuable work in the future.

ACKNOWLEDGEMENTS

Our Sincere thanks to our Professor Dr.S.V.Saravanan, Principal, Christ the King Engineering College for his valuable support in documenting and framing a new approach in this paper from various previously published articles in the journal.

REFERENCES

[1] R. R. Korfhage. Information Storage and Retrieval. Willey Computer Publishing, 1997.

[2] G. Salton and M. J. McGill. Introduction to Modern Information Retrieval. McGraw–Gill, 1983.

[3] R. B. Yates and B. R. Neto. Modern Information Retrieval. Pearson Education, 1999.

[4] G. W. Cottrell B. T. Bartel and R. K. Belew. “Learning to retrieve information”, In Current Trends in Connectionism: Proceedings of the Swedish Conference on Connectionism, pp. 345–354, 1995.

[5] C. J. Van Rijsbergen. Information Retrieval. Butterworth-Heinemann, 1979.

[6] E. A. Fox and J. A. Shaw.” Combination of multiple searches”. In Proceedings of the Second Text Retrieval Conference TREC 2, pp. 243–252,1994.

[7] M. Montague and J. A. Aslam,” Condorcet fusion for improved retrieval”,.In Proceedings of the `

[8] G. Mauris L. Valet and P. Bolon ”A statistical overview of recent literature in information fusion”, In Procedings of the Third International Conference on Information Fusion, pp. 22–29, July 2000.

[9] Joon Ho Lee,” Combining multiple evidence from different properties of weighting schemes”, In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 180–188, 1995.

[10] Joon Ho Lee,” Analyses of multiple evidence combination”, In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 260–276, 1997.

[11] Joon Ho Lee,” Combining multiple evidence from different relevant feedback networks”, In Proceedings of the 5th international Conference on Database Systems for Advanced Applications, pp. 421–430, 1997.

[12] Savoy. ,Combining probabilistic and vector schemes,”In Proceedings of the Fourth Text Retrieval Conference, pp. 537–548, 1996.

[13] C. A. Coelo Coelo. “An updated survey of ga-based multiobjective optimization techniques”, ACM Computing Survey, pp. 109–143, Jan 2000.

[14] H. Bilhart. “Learning retrieval expert combinations with genetic algorithm”, International Journal of Uncertainity,Fuzziness and Knowledge Based Systems, 11 (1):87–114, Feb 2003.Kevin R. Fall, W. Richard Stevens, TCP/IP Illustrated, Volume 1: The Protocols, 2nd ed., Addison-Wesley, USA, 2011.

[15] Jian Zhang, Jianfeng Gao, Ming Zhou, Jiaxing Wang,`` Improving the Effectiveness of Information Retrieval with clustering and Fusion”, To appear in the Computational Linguistics and Chinese Language Processing,

[16] Rabia Nuray and Fazli Can. Automatic ranking of information retrieval systems using data fusion. Information Processing and Management, 42(3):595–614, May 2006.