© 2016 IJSRSET | Volume 2 | Issue 2 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Expected Loss Optimization for Document Ranking by Active Learning

G Saranya*, M Manikandan

Department of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, Tamil Nadu, India

ABSTRACT

Learning to rank is the emerging research field in many data mining applications and information retrieval techniques (e.g. Search engines). The major issue in ranking algorithm is that the quality or ranking is affected by labeled examples, since it is very expensive and also time consuming to collect labeled samples. This problem brings a great need for active learning algorithm; however, in literature learning to rank uses supervised learning algorithm where ranking is based on labeled data only. A general active learning framework Balanced two stage Expected Loss Optimization is proposed to select the most informative document based on user's query. The algorithm is based on two levels, Query level and Document level and grade distribution is done based on query and document pairs. Experiment on web search dataset has demonstrated with the proposed algorithm.

Keywords : Expected Loss Optimization, Active Learning, Ranking And Supervised Learning

I. INTRODUCTION

Data mining can extract the useful and important information from a large set of database. The other important techniques in data mining related to extracting useful information are Text mining, Pattern mining, Information extraction, Information Retrieval etc. It is very difficult for a user to find high quality of document when there are many documents related to his search. User only wants the documents to be listed related to his query. Ranking is a core problem in many information retrieval systems. Modern search engines[8] such as Google, yahoo, Bing, ask, especially those search engines designed for the World Wide Web, commonly analyze and combine hundreds of features extracted from the submitted query and all related documents. The ultimate goal of ranking [9] is that given a query, the documents has to be ranked based on the maximum occurrence of the query term in the document. The sheer size of information available in World Wide Web which leads to the great need of ranking system to get only the most informative and relevance document instead of displaying the list of documents which is not useful for the user. Also the users are interested only in the top ranked documents which contain more information related to the given query.

Active learning is also known as query learning and it is subfield for machine learning and artificial intelligence. The basic hypotheses behind active learning algorithm are that it can choose the data from where it learns. Active learning algorithms [1] are well-motivated in many modern machine learning problems where the data may be abundant but it reduces the labeling effort than compared with many machine learning algorithms. In many other supervised learning algorithms the quality of the ranking is affected with the labeled data which contains irrelevant documents matching the query. Compared with the active learning for classification, active learning for ranking faces some of the unique challenges such as there is no notion for classification margin in ranking function.

A. Related Work

A wide variety of ranking algorithm for ranking document based on query has been proposed in the literature. Unfortunately, there is neither a standard problem definition nor a standard ranking algorithm is proposed yet. Most existing algorithm are based on Pairwise[3], List wise[5] and Point wise approach which attempts to rank the documents based on comparison with other document.

KhaledAlsabti and Sanjay Ranka [14] proposed an algorithm Dynamic Page Rank, which identifies all the query words in the document. The query word can be enhanced by tokenization, Stemming, stop words removal as well as sense disambiguation approach. The resulted queries are passed to the search engine where the documents are retrieved based on the enhanced query. The web pages are ranked from higher to lower dynamic page rank values. The user receives more meaningful contents at top of the search results where least preferred documents are displayed at the last position in rank list.

In 2011 Martin Szummer [2] proposed semi-supervised ranking algorithm, the algorithm tends to include non-informative documents when there are more number of documents associated with each query. Query efficient algorithm [3] was proposed by Weiss Y. and Torralba.. The algorithm is based on pair wise preferences with optimal query selection. It cannot find optimal solutions in the case of larger query in document selection.

Ali Mohammad ZarehBidoki and Nasser Yazdani proposed a novel Distance Rank algorithm [15] based on recursive method. The algorithm is based on the distance factor where the distance between the web pages are calculated to compute the rank in search engine. The main advantage of this algorithm is, it can find the pages faster and more quickly based on the distances solution. The distance rank algorithm adopts some properties of page rank. The page rank will have high rank value if it have more incoming link on a page.

In 2008 Snelson.E [7] proposed soft rank algorithm and Gaussian processes. The similarity score for the documents are random so there is a possibility of ranking the document at any position in rank list. AdaRank [16] an boosting algorithm which is proposed by Xu and Li. Boosting is a general technique for improving the ranking performance and also it offers many advantages like easy implementation, efficiency in retrieval of relevant document and also accuracy in ranking.

Khaled Alsabti and Sanjay Ranka [6] proposed dynamic page rank algorithm, which identifies all query words in document and appropriate sense is assigned to each occurrence of word in textual context. The query terms are separated by applying various steps like tokenization, stemming, stop word removal as well as sense disambiguation approach.

M. ZarehBidoki and N. Yazdani [14] have analyzed that the World Wide Web, he proposed an algorithm Distance Rank algorithm which is based on Visits of Links (VOL) being devised for search engine. This algorithm is proposed to display the web pages based on maximum visits that is the web page which is visited by many user will be displayed on the top position in rank list. So each and every page is given a page weight according to the visits of the user. Based on the literature study the Dynamic Page Rank and Weighted Page Rank shares the common ranking functionality of basic Page Rank algorithm [15]. Google Search engine uses Page Rank algorithm which is efficient and low cost.

B. Datasets

The type of dataset used for this experiment is web search dataset from a commercial search engine. The data set consists of a random sample of about 100 documents. The real time datasets are collected from internet. Each document may contain the tags such as number, title, description and narrative. Under those tags the details related to the documents are represented. The query-document pairs are labeled using a five-grade labeling scheme: {Bad, Fair, Good, Excellent and Perfect}.

Query features, dependent on the query only and have constant values across all the documents, for example, the query should not be a person name or place name. Document features, dependent on the documents only and have constant values across the datasets. Query-document features, dependent on the relation of the query with respect to the document, for example, the number of times each term in the query appears in the document.

C. Modules And Description

The modules involved in this experiment are as follows:

i. Data Selection and POS Tagging

In this module, input dataset is selected from the directory. part-of-speech tagging (POS tagging or POST), also called grammatical tagging of words or word-category disambiguation, is the process of marking up a word in a text and tag the words according to particular part of speech. For e.g. If there is a word like person name Ram then it will be tagged as <NN>Ram representing as a noun for easy text classification.

ii. Data Preprocessing

Preprocessing is one of the data filtering technique to reduce the garbage values from dataset. The Preprocessing process will remove unwanted tags from the datasets and after the removal for stop words the remains words are taken as query which is considered as meaningful words.

iii. Compute TF, IDF, TFIDF

TF and IDF is the short form of Term FrequencyInverse Document Frequency. It is a numerical statistic which intended to reflect how important a word is to a document in a collection dataset. It is often used as a weighting factor in information retrieval and text mining. The similarities between the documents are calculated using cosine similarity by considering the values of Term Frequency and Inverse Document Frequency.

iv. Compute Correlation and Clustering.

The correlation is used to find the relationship between one document and another the values may ranges from +1 and −1, where 1 is total positive correlation, 0 is no correlation, and −1 is total negative correlation. After finding the correlation between documents, clustering is performed based on correlation value.

v. Ranking documents

In this module, ranking the documents is performed based on the occurrence of query term in each document. Based on user query, related documents are retrieved and ranked in perfect position.

II. METHODS AND MATERIAL

Two stage ELO (Expected Loss Optimization) [1] algorithm which uses term frequency to select most informative examples that minimizes loss during document selection. First stage in ELO is used in query selection and second stage is document selection.

The input instance is a query and a set of documents associated with it, while the output is a vector of relevance scores. Based on the relevance score document ranking is done through the repetition of that particular query term in the documents if the query term is found more in an document then it is ranked in first position in ranking list. If the query term is repeated very less in a document then it will be in last position in ranking list. Thus according to the query the user will gain information. Expected loss optimization gives importance for both query and document level which improves the ranking performances and also it gives the user the most informative examples and relevant document with respect to the query.

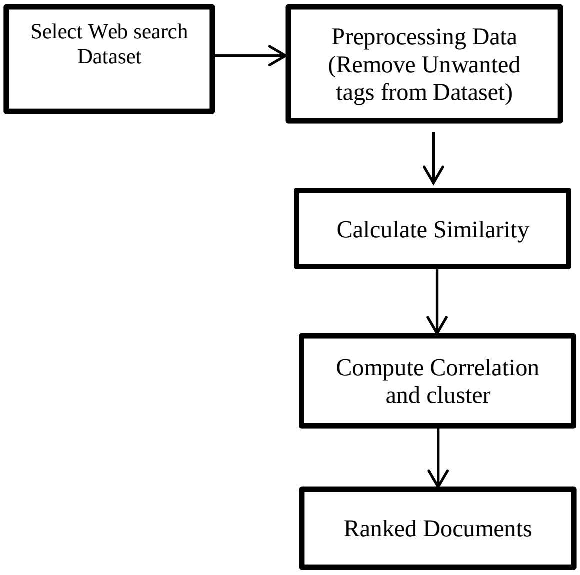
Select Web search Dataset
Preprocessing Data (Remove Unwanted tags from Dataset)
Calculate Similarity
Compute Correlation and cluster
Ranked Documents

Figure 1. System Design

A. Query Level

In query level, the input is set of documents and from all the documents the meaningful word is selected after the removal of stop words. After that process frequent occurrence of those words (query) in each document is calculated by Term Frequency (TF).

$$\text{TF (t)} = \frac{\text{Number of times term t appears in a document}}{\text{Total number of unique terms in a document}} \quad \text{-(1)}$$

Inverse Document Frequency (IDF), the number of documents containing the particular query.

$$\text{IDF (t)} = \log \frac{\text{Total number of documents}}{\text{Number of documents with term t}} \quad \text{------ (2)}$$

Similarity between the documents is calculated by considering both TF and IDF. If the query occurs in more than one document, based on that the similarity score is considered. If the particular query occurs only in one document then it will be discarded in query list.

B. Document Level

In document level, the correlation between each and every document is calculated i.e., the relationship between one document and another. The correlation between the documents is calculated as follows,

$$\text{Cos}\ (d_i\ , d_j) = d_i . d_j \qquad \text{-----------}\ (3)$$

Where d_i and d_j represents the documents. If the Cos value is 0 then, there is no similar terms between documents. The value of Cos is 1 or positive values the documents may contain similar query terms. Based on correlation value the documents can be clustered, similar values are grouped in one cluster. Each and every cluster is given a cluster ID. For example, if a query is repeated in more than one document then the documents are clustered for easy retrieval of the document. In document level the input instances is a set of query the user can select the query from list and set of document will be displayed associated with the query.

III. RESULTS AND DISCUSSION

We implemented the algorithm using java. A Java program runs exactly the same way on all computers. Most other languages allow small differences in interpretation of the standards.

The input to the experiment is set of documents (datasets). Query terms are separated during query level processing and after getting the meaning words from the document set, the documents can be retrieved using those words. Finally ranking process is done by considering the maximum query occurrence in a document. Grade distribution for the existing system considered only based on ranking. If the document is rated as perfect then ranking position of the document is high. The computation time is also calculated for both existing system and proposed system. The comparison between the execution time is shown in Fig. 3. Where the proposed system requires less time to compute informative queries and rank the documents.

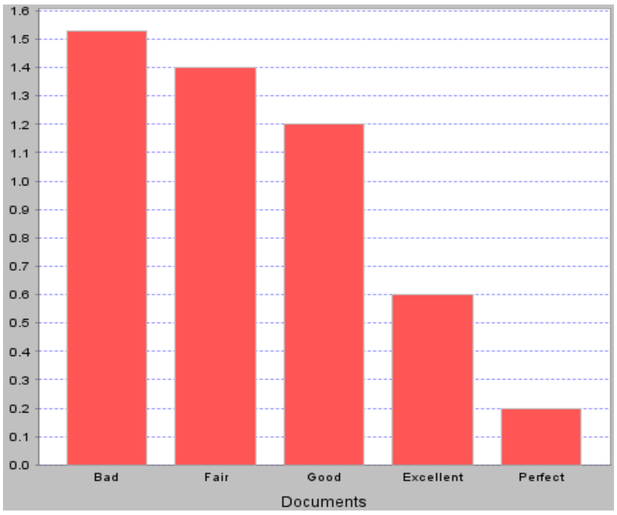

Figure 2. Grade Distribution

Fig 2 represents the grade distribution for query and document selection. The documents can be rated according to the occurrence of the query term. If the document contains the query only once then it will be rated as bad such that if a document has maximum query terms then it will be rated as perfect.

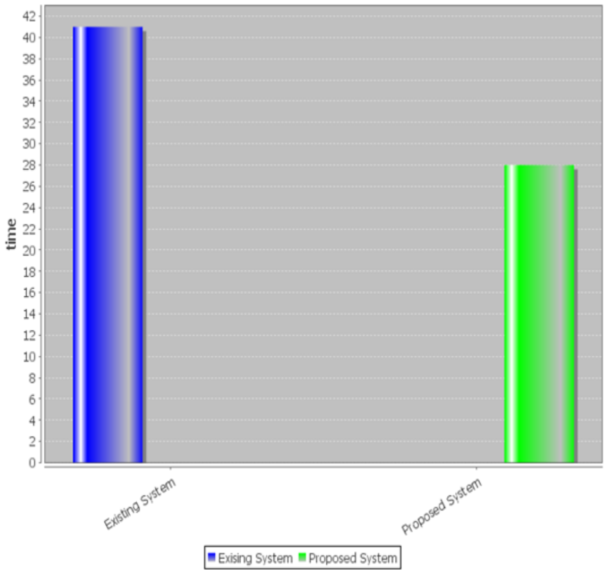

Figure 3. Comparison graph

The above graph shows the comparison between the working of the existing algorithm and proposed algorithm. The execution time taken for the existing algorithm to compute the query term and ranking is more than the active learning approach using Expected Loss Optimization

IV. CONCLUSION

As information becomes available in digital form, people expect to use this information effectively. The goal of any information retrieval system is to retrieve documents that match the information need of the user. But it is very difficult to specify the information need to the IR system. Sometimes, even the users do not know their information need precisely. To calculate the relevance, existing IR systems considered features only from the document itself. Active learning approach is used to select the meaning words from the document which is referred to as query and rank the documents based on selected query terms. This active learning approach was best suitable for retrieval of informative document from well-controlled repositories like web search data, where the user can find their relevance and most informative document by ranking in perfect position in rank list.

V. REFERENCES

[1] Bo Long and Jiang Bian, "Active Learning for Ranking through Expected Loss Optimization", IEEE Transactions on Knowledge and Data Engineering, Vol. 27, No.5, 2015.

[2] Martin Szummer, "Semi-supervised learning to Rank with Preference Regularization", in Proc. 32nd Int ACM SIGIR Conf. Res. Develop. Inform. Retrieval,pp. 662–663, 2011.

[3] Qian B. Li H. Wang J. Wang X. and Davidson I, "Active Learning to Rank using Pairwise Supervision", Proc. 13th SIAM Int. Conf. Data Mining, pp. 297–305, 2013.

[4] Wenpu Xing and Ghorbani Ali, "Weighted Page Rank", Proceedings of the IEEE International Conference on Computer Science, 2004.

[5] Xia F. Liu T. Wang J. Zhang W. and Li H, "List wise approach to learning to rank: theory and algorithm", Proc. 25th Int. Conf. Mach. Learn., pp. 1192–1199, 2008.

[6] Khaled Alsabti Sanjay Ranka and Vineet Singh,"Efficient Information Retrieval Using Dynamic Page Rank Algorithm", In Proceedings of IPPS/SPDP Workshop on High Performance Data Mining, 2011.

[7] Guiver J. and Snelson E, ‘Learning to rank with SoftRank and Gaussian processes’, Proc. 31st Ann. Int. ACM SIGIR Conf. Res.Develop. Inform. Retrieval,pp. 259–266, 2008.

[8] R.Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval: The Concepts and Technology behind Search, 2nd Addison-Wesley Professional, 2011.

[9] C. D. Manning, P. Raghavan and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[10] L.Yang, L. Wang, B. Geng, and X.-S.Hua. "Query sampling for ranking learning in web search". In SIGIR'09: Proceedings of the 32nd international ACMSIGIR conference on Research and development in information retrieval, pages 754-755, New York, NY, USA, 2009.

[11] C.Campbell, N. Cristianini, and A. Smola. "Query learning with large margin classifiers". In Proceedings of the Seventeenth International Conference on Machine Learning, pages 111-118. Morgan Kaufmann, 2000.

[12] D. Cossock and T. Zhang, "Subset ranking using regression" .In Proc. Conf. on Learning Theory, 2006.

[13] E. Yilmaz and S. Robertson,"Deep versus shallow judgments in learning to rank". In SIGIR '09:Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pages 662-663, New York, NY, USA, 2009.

[14] A. M. ZarehBidoki and N. Yazdani, "DistanceRank: An intelligent ranking algorithm for webpages" information Processing and Management, Vol 44, No. 2, pp. 877-892, 2008.

[15] Rekha Jain, DrG.N.Purohit, "Page Ranking Algorithms for Web Mining", International Journal of Computer application,Vol 13, Jan 2011.

[16] Xu J. and Li H., "AdaRank: A boosting algorithm for information retrieval", SIGIR, pp. 391–398, 2007