ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

A Small Analysis on Learning to Rank for Information Retrieval

G Saranya[1], G Swetha[2]

P.G Scholar, Dept of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, India[1]

Assistant Professor, Dept of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, India[2]

Abstract: Learning to rank for information retrieval has gained a lot of interest in the recent years because, ranking is the central problem in many information retrieval applications, such as document retrieval, collaborative filtering, question answering, multimedia retrieval, text summarization, and online advertising machine translation etc. The extremely large size of the Web documents makes it generally impossible for the common users to find their desired information by surfing the Web. As a consequence, effective and efficient information retrieval has become more important and also search engine (information retrieval system) has become an essential tool for people to locate their needed information. So, we propose novel active learning algorithm that is two stages Expected loss optimization (ELO), which minimizes the expected loss of information and rank the document which is more relevant to the query and gives the user the most informative document instead of displaying all the related documents which is not useful for the user.

Keywords: Ranking, document, query, Expected loss Optimization, Learning to rank.

I. INTRODUCTION

Learning to rank is a relatively new research area which has emerged in the past decade [19]. Search engines are essential tools for finding and exploring information on the Web and other information systems. To a larger extent the quality of a search engine is determined by the ranking function used to produce the results according to user's query. Ranking is the core of information retrieval system; user gives the query the documents have to be ranked according to their relevance to the query. Machine learning algorithms are used to learn the ranking function. Ranking has widespread applications such as commercial search engines and recommend system, which can find out relevance between the relevant documents in context of given user's query, and place them in order of their relevance in rank list. In classification, the queries and documents are given; each query is associated with a Perfect ranking list of the documents. The ranking model is then created using the classification process according to the given query. In contrast, learning to rank approaches [21] [2] in information retrieval allows retrieval systems to incorporate hundreds or even thousands of arbitrarily defined features. Most importantly, these approaches automatically learn the most effective combination of these features in the ranking function based on the available data for classification. Some evaluation metrics are needed to measure the quality of search engine; one of the most commonly used metric in web search ranking is Discounted Cumulative Gain (DCG), the Discounted Cumulative Gain is used to measure ranking quality of search engines. In information retrieval, it is often used to measure effectiveness of web search engine algorithms or other related applications.DCG measures the usefulness or gain, of a document based on its perfect position in the rank list. If the relevant document is in lower position then it is less useful for the user to gain knowledge. The purpose of the paper is to integrating both query level selection and document level selection for ranking and we proposed an expected discounted cumulative gain (DCG) loss optimization (ELO-DCG) algorithm, to select most informative and relevant document associated to query.

This paper is organized as follows. In section 2 related works is discussed. In section 3 we present the evaluation methodology of active learning to rank and ELO for raking. The review question and observations are presented in section 4 and section 5.In section 6 discussions and suggestion are presented. Finally the drawbacks and conclusion is presented in section 7.

II. RELATED WORK

Learning to rank [20] has three common approaches they are: Point wise approaches, Pair wise approaches, and List wise approaches. These three different approaches learn to rank in different ways. That is, they may define different input and output spaces, use different hypotheses, and employs different loss functions.

The Point wise approaches are the earliest approaches,[2] the basic hypotheses of this approach is to map the document's ordinal scales into numeric values using regression and classification method, it try to compare the relevance score of every two documents, then comparison result is produced. Based on that result the document will be ranked.

Pair wise approach uses binary classifier method that will tell which document is better in a given pair of documents. The goal of using binary classifier is to minimize average number of inversions in ranking functions.

DOI 10.17148/IJARCCE.2015.4931

The List wise approaches [21] is similar with the basic idea of pair wise approach, it directly compare the relevance list of documents based on query, instead of trying to get ranking score for each document individually. It uses Ad a Rank and soft Rank algorithms for ranking. Compared with traditional active learning algorithm; there is still limited work on the active learning for ranking in recent years.

The problem of document selection based on query in ranking is studied by Donmez and Carbonell [3].The uncertainty sampling is simple and common strategy in active learning, the issue in sampling is that the algorithm selects queries for which the label uncertainty samples have highest relevance score [4].The main drawback in this type of approach is noise and variance. Active learning algorithm minimizes the noise and reduces variance proposed in [5].

Query by Committee algorithm [8] uses noise free classification function. Another common approach for active learning is to select query that once added to training set which leads to large increase in the objective function value that is being optimized [6].

Many other ranking algorithms such as Rank SVM [11] and Rank Boost [7] suggests to add the most relevant pairs of documents to the training set, the document's predicted relevance scores are very close under the current ranking models. In the term of binary relevance, greedy algorithm [10] is proposed which selects the document that differentiates two different ranking systems in terms of average precision. The comparison of effective and efficient document selection methodologies in learning to rank are found in [15].

L. Yang, L. Wang [12] proposed greedy query selection algorithm that minimizes query density and query diversity. Some empirical and theoretical work related to query sampling are found in[13] the results shows that better having more queries but less number of documents per query than having more documents and less queries.

III. EVALUATION METHODOLOGY

A. Active Learning to Rank

Active learning for ranking reduces the labeling effort than compared to supervised learning. In many other supervised learning algorithms the quality of the ranking is affected with the labeled data which contains irrelevant documents matching the query. Existing approaches such as[21][2] for ranking are not readily applicable to rank. Compared with the active learning for classification [3], active learning for ranking faces some of the unique challenges such as there is no notion for classification margin in ranking function. Some active learning approach like [14] Query by Committee (QBC) has not justified for ranking under regression and classification framework [15]. Active learning for ranking can select examples at different levels, one is query level and other is document level. Query level selects informative queries with all associated documents. Document level selects each and every document individually for a given query. Since query level and document level active learning has its own drawback [16], an informative query could be missed if none of its documents is selected, or only one document is selected per query, which is not a good example in learning to rank.

B. Two Stage ELO algorithm

Active learning framework, expected loss optimization (ELO) for both query and document selection is applied for ranking. There is a great need for active learning framework when selecting the data for ranking. The basic idea behind the proposed algorithm is that given a loss function, the samples minimizing the expected loss (EL) which are considered as most informative document. Two stage ELO algorithm which uses function ensemble to select most informative examples that minimizes a chosen loss.

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 the 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 reduces the discounted cumulative gain (DCG) loss, which is the main problem in ranking systems for information retrieval.

IV. RESEARCH QUESTIONS

A research questions plays a major role in the survey and it provides clarity for the survey. The questions related to information retrieval and ranking are described as follows.

Q1. What is the purpose of information Retrieval (IR) System?
Motivation: To analyze why we use information retrieval system and its advantage
Q2. How Information Retrieval Systems Works?
Motivation: To know the working of Information retrieval system.
Q3. What are the important problems in information retrieval?
Motivation: Problems in information retrieval are considered
Q4. What are terms considered during ranking the document?
Motivation: To know the performance of ranking.
Q5. How ranking is done in Search engine?
Motivation: To analyze the process of ranking in search engine

V. OBSERVATION

Based on the research about ranking in information retrieval the questions arises are as follows and the

DOI 10.17148/IJARCCE.2015.4931

answers for research question are observed and presented below in terms of retrieval process

Q1. What is the purpose of information Retrieval (IR) System?

Information Retrieval (IR) is all about the process of providing answers to user as per the information they need. It is concerned with the collection, representation, storage, accessing, manipulation and display of the information necessary to satisfying user's needs. IR system is to provide information that changes the knowledge state of a user so that the user can able to perform a present task and also prepared to perform future tasks in other ways to lead a better quality of life.

Q2. How Information Retrieval Systems Works?

IR is a component of an information system. An information system must make sure that everybody meant to be served with the information they needed to accomplish their tasks, solve problems, and make their own decisions, but no matter where that information is available to them.

An information system must (1) actively find out what users needs, (2) acquire documents (like programs, or data items, or products), resulting in a collection, and (3) Relatively match documents with user's needs.

Q3. What are the important problems in information retrieval?

Some of the problems in information retrieval system are, the human-computer interface, knowledge representation, Procedures for processing knowledge/information, Designing user-enhanced information systems and System evaluation

Q4. What are terms considered during ranking the document?

The important things need to be considered during ranking the document are as follows:1) Term importance 2)Stemming 3)Query expansion 4)Document structure 5)Personalization

Term importance: Frequent (repeated) vs. discriminative words are important when ranking a document.

Stemming:Stemming is the process of morphologically equivalent words (e.g. bicycles → bicycle)

Query expansion: Query expansion relates, which are semantically equivalent it is similar to stemming process (e.g. bicycles → bicycle)

Year	**Title**	**Methodology/ Algorithm**	**Inference**
2003	An Efficient Boosting Algorithm for Combining Preferences	Rank Boost	Most relevant information to be combined represents relative preferences rather than absolute rating and ranking
2006	Minimal Test Collections for Retrieval Evaluation	link evaluation with test collection	Obtaining labelled examples for data is very expensive and also time-consuming is high
2008	Optimizing Estimated Loss Reduction for Active Sampling in Rank Learning	SVM-based and boosting-based rank learning	It does not distinguish between the relative order of two relevant or two non-relevant examples
2008	Active Preference Learning with Discrete Choice Data	active learning algorithm	This algorithm is not possible to evaluate over the entire document ranking because of labelled data
2011	Semi-supervised learning to Rank with Preference Regularization	Semi-supervised ranking algorithm.	It tend to include non-informative documents when there are a large number of documents associated with each query
2012	An Active Learning Algorithm for Ranking from Pairwise Preferences with an Almost Optimal Query Complexity	Query Efficient Algorithm	Algorithm cannot be used to find almost optimal solutions in case of larger query in document selection

TABLE 1: LIST OF RANKING ALGORITHM BASED PAPERS

Document structure: Matching of query terms in different parts of the document is important (e.g. title, body description etc)

Personalization: We can also consider user's information to improve ranking functionality.

Q5. How ranking is done in Search engine?

Search engines rank the web pages [18] by their expected relevance to a user's query based on two different

DOI 10.17148/IJARCCE.2015.4931

methods they are: Query-dependent and Query-independent methods.

Query-independent method is used to measure the estimated importance of a page, independent consideration of how well the page matches with the specific query. Query-independent ranking is usually based on link analysis method, for examples it includes Page Rank and also Trust Rank.

Query dependent methods attempt to measure the degree of page which matches to a specific query; independent importance to the page is given. Query- dependent ranking is based on heuristics mechanism that usually consider the location and number of matches to various query words in the document on the page itself or in any anchor text referring to the particular page or in any URL, for example Boolean model, vector space model etc.

VI. DISCUSSION AND SUGGESTION

In recent days the necessary and importance of learning to rank is increased in all the fields, here we focus only on the information retrieval. So searching and retrieving the relevant document from the large sets of data is becoming a critical task. To overcome this problem, active learning for ranking techniques was followed. Since ranking is core component in information retrieval system where the user only interested in top listed document instead of displaying all the document, which are unnecessary to the user.

Ranking are used in many other fields such as Mobile application to rank and rate the performance of the each and every application, In Educational field to evaluate the performance of the Student and to award the best student in the academic year, industries uses ranking measures for their employees to know their activities and to improve their performance in job. Raking is used in sports to rank the performance of the player. Now a day's ranking are used to know about the details of the organizations such as colleges, schools etc. People are interested to know the top colleges and schools to educate their children. Learning to rank have to be developed more in other fields also. It is recommended to implement the ranking system all over world for the best knowledge for the users to acquire their needed information in correct time.

VII. CONCLUSION

As technology improves everyday new developments are constantly infiltrating our lives. The research in learning to rank is an outgoing process and the requirement of ranking change every day based on the requirements from the user. Active learning for ranking is differs from Active learning for classification and regression, in addition active learning for ranking has some unique features. In literature there are many ranking algorithm they are all time consuming and also cost much in obtaining labeled data compared with those algorithm Expected loss optimization for query and document level ranking by active learning performs efficiently by providing the user the most informative documents for their references.

REFERENCES

[1] Bo Long, Jiang Bian, Olivier Chapelle, Ya Zhang, Yoshiyuki Inagaki, and Yi Chang "Active Learning for Ranking through Expected Loss Optimization. VOL. 27, NO. 5, MAY 2015.

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

[3] P. Donmez and J. G. Carbonell. Optimizing estimated loss reduction for active sampling in rank learning. In ICML '08: Proceedings of the 25th internationalconference on Machine learning, pages 248-255, New York, NY, USA, 2008. ACM.

[4]D. Lewis and W. Gale. Training text classifiers by uncertainty sampling. In Proceedings of the 17thAnnual International ACM SIGIR Conference onResearch and Development in Information Retrieval,pages 3-12, 1994.

[5] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. In G. Tesauro, D. Touretzky, and T. Leen, editors, Advances inNeural Information Processing Systems, volume 7,pages 705-712. The MIT Press, 1995.

[6] 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.

[7]Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. Anefficient boosting algorithm for combining preferences.Journal of Machine Learning Research, 4:933-969,2003.

[8] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby.Selective sampling using the query by committeevalgorithm. Machine Learning, 28(2-3):133-168, 1997.

[9] A. Aslam, E. Kanoulas, V. Pavlu, S. Savev, and E. Yilmaz,"Document selection methodologies for efficient and effective learning-to-rank," in Proc. 32nd Int. ACM SIGIR Conf. Res. Develop.Inform. Retrieval, 2009, pp. 468–475.

[10] B. Carterette, J. Allan, and R. Sitaraman.Minimal test collections for retrieval evaluation.In Proceedings of the 29th annual international ACM SIGIRconference on Research and development ininformation retrieval.ACM, 2006.

[11] R. Herbrich, T. Graepel, and K. Obermayer.Large margin rank boundaries for ordinal regression. In Smola, Bartlett, Schoelkopf, and Schuurmans, editors,Advances in Large Margin Classifiers. MIT Press,Cambridge, MA, 2000.

[12] 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 ininformation retrieval, pages 754-755, New York, NY,USA, 2009. ACM.

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

[14] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby.Selective sampling using the query by committee algorithm. Machine Learning, 28(2-3):133-168, 1997.

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

[16] M. E. Maron, J. L. Kuhns, and L. C. Ray, 'Probabilistic indexing: A statistical technique for document identification and retrieval', Thompson Ramo Wooldridge Inc, Los Angeles, California, Data Systems Project Office, Technical Memorandum 3, Jun. 1959.

[17] K. Spärck Jones, Ed., Information Retrieval Experiment. Butterworth-Heinemann, 1981

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

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

[20] LI, Hang. "A Short Introduction to Learning toRank."2011.<http://research.microsoft.com/en-us/people/hangli/l2r.pdf>.

[21] F. Xia, T.-Y. Liu, J. Wang, W. Zhang, and H. Li, "Listwise approach to learning to rank: theory and algorithm," in Proc. 25th Int. Conf. Mach. Learn., 2008,pp.1192–11.

DOI 10.17148/IJARCCE.2015.4931