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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

Active Learning for Ranking Through Support Vector Machine

G.Shalini[1], P. Rethina Sabapathi[2]

PG Scholar, Dept. of CSE, Sri Venkateswara College of Engg and Tech, Thiruvallur, India[1]

Associate Prof, Dept. of CSE, Sri Venkateswara College of Engg and Tech, Thiruvallur, India[2]

ABSTRACT: Learning to rank arises in many data mining applications, ranging from web search engine, online advertising to recommendation system. In learning to rank, the performance of a ranking model is strongly affected by the number of labeled examples in the training set on the other hand, obtaining labeled examples for training data is very expensive and time-consuming. This presents a great need for the active learning approaches to select most informative examples for ranking learning; however, in the literature there is still very limited work to address active learning for ranking. In this paper, we propose a general active learning framework, expected loss optimization (ELO), for ranking. The ELO framework is applicable to a wide range of ranking functions. Under this framework, we derive a novel algorithm, expected discounted cumulative gain (DCG) loss optimization (ELO-DCG), to select most informative examples. Then, we investigate both query and document level active learning for raking and propose two-stage ELO-DCG algorithm which incorporate both query and document selection into active learning. Furthermore, we show that it is flexible for the algorithm to deal with the skewed grade distribution problem with the modification of the loss function. Extensive experiments on real-world web search data sets have demonstrated great potential and effectiveness of the proposedframework and algorithms.

I. INTRODUCTION

Ranking is the core component of many important information retrieval problems, such as web search, recommendation, computational advertising. Learning to rank represents an important class of supervised machine learning tasks with the goal of automatically constructing ranking functions from training data. As many other supervised machine learning problems, the quality of a ranking function is highly correlated with the amount of labeled data used to train the function. Due to the complexity of many ranking problems, a large amount of labeled training examples is usually required to learn a high quality ranking function. However, in most applications, while it is easy to collect unlabeled samples, it is very expensive and consuming to label the samples. Active learning comes as a paradigm to reduce the labeling effort in supervised learning. It has been mostly studied in the context of classification tasks [20]. Existing algorithms for learning to rank may be categorized into three groups: point wise approach, pair wise approach, and list wise approach. Compared to active learning for classification, active learning for ranking faces some unique challenges. First, there is no notion of classification margin in ranking. Hence, many of the margin-based active learning algorithms proposed for classification tasks are not readily applicable to ranking. Furthermore, even some straightforward active learning approach, such as query-by-committee (QBC), has not been justified for the ranking tasks under regression framework. Second, in most supervised learning
setting, each data sample can be treated completely independent of each other. In learning to rank, data examples are not independent, though they are conditionally independent given a query. We need to consider this data dependence in selecting data and tailor active learning algorithms according to the underlying learning to rank schemes. Third, ranking problems are often associated with very skewed data distributions. For example, in the case of document retrieval, the number of irrelevant documents is of orders of magnitude more than that of relevant documents in training data. It is desirable to consider the data skewness when selecting data for ranking. Therefore, there is a great need for an active learning framework for ranking. In this paper, we attempt to address those three important and challenging aspects of active learning for ranking. We first propose a general active learning framework, expected loss optimization (ELO), and apply it to ranking. The key idea of the proposed framework is that given a loss function, the samples minimizing the expected loss (EL) are the most informative ones.

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

Under this framework, we derive a novel active learning algorithm for ranking, which uses function ensemble to select most informative examples that minimizes a chosen loss. For the rest of the paper, we use web search ranking problem as an example to illustrate the ideas and perform evaluation. But the proposed method is generic and may be applicable to all ranking applications. In the case of web search ranking, we minimize the expected discounted cumulative gain (DCG) loss, one of the most commonly used loss for web search ranking. This algorithm may be easily adapted to other ranking loss such as normalized discounted cumulative gain (NDCG) or average precision. To address the data dependency issue, the proposed algorithm is further extended to a two-stage active learning schema to seamlessly integrate query level and document level data selection. Finally, we extended the proposed algorithms to address the skew grade distribution problem.

The main contributions of the paper are summarized as follows.

We propose a general active learning framework based on expected loss optimization. This framework is applicable to various ranking scenarios with a wide spectrum of learners and loss functions. We also provide a theoretically justified measure of the in formativeness.

Under the ELO framework, we derive novel algorithms to select most informative examples by optimizing the expected DCG loss. Those selected examples represent the ones that the current ranking model is most uncertain about and they may lead to a large DCG loss if predicted incorrectly.

We propose a two stage active learning algorithm for ranking, which addresses the sample dependence issue by first performing query level selection and then document level selection.

We further show how to extend ELO framework to address the skewed grade distribution problem in ranking. Balanced version ELO algorithms are derived for both query level active learning and two stage active learning.

II. EXISTING SYSTEM

Learning to rank represents an important class of supervised machine learning tasks with the goal of automatically constructing ranking functions from training data. As many other supervised machine learning problems, the quality of a ranking function is highly correlated with the amount of labeled data used to train the function. Due to the complexity of many ranking problems, a large amount of labeled training examples is usually required to learn a high quality ranking function. However, in most applications, while it is easy to collect unlabeled samples, it is very expensive and time-consuming to label the samples.

DISADVANTAGE OF EXISTING SYSTEM:

- Less attention is paid to the privilege control and the identity privacy.
- Collusion attack.

III. PROPOSED SYSTEM

First, there is no notion of classification margin in ranking. Hence, many of the margin-based active learning algorithms proposed for classification tasks are not readily applicable to ranking. Furthermore, even some straight forward active learning approach, such as query-by-committee, has not been justified for the ranking tasks under regression framework. Second, in most supervised learning setting, each data sample can be treated completely independent of each other. In learning to rank, data examples are not independent, though they are conditionally independent

ADVANTAGE:

- Flexible.
- On-Demand.
- Low-Cost usage of computing resources.

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

IV. SYSTEM ARCHITECTURE

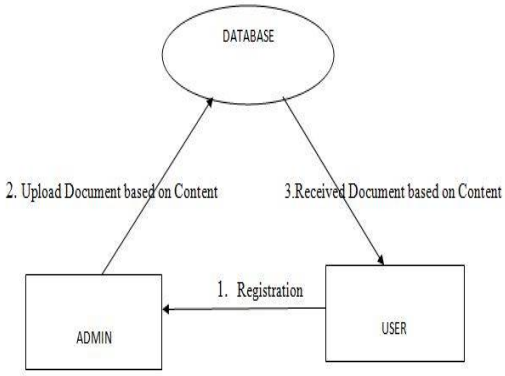

Fig.1 System Architecture

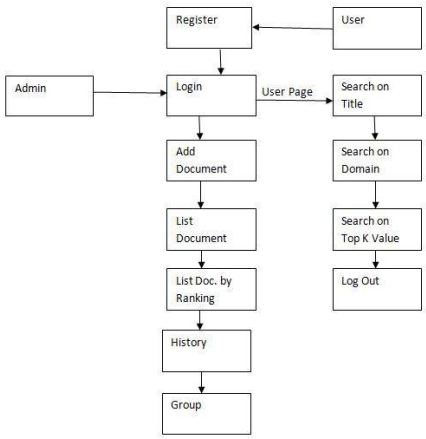

Fig. 2 Data Flow Diagram

MODULES:

- Admin
- User.

Module 1: Admin.

The Admin should login with database to upload file based on the content and data with their allotted size and within the period of time. They can view the uploaded document list and also view based on the Ranking list

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

documents. Admin will have the permission to see the downloaded history's such as file id and user id and then date. Here graph are generated for the expectation measurement based on the matching document and loss document.

Module 2: User.

The Consumers should register with server and then login. Here consumer can search the document based on various contents techniques like as ranking method. They can search the files based on the title and based on the domain then finally based on the top k value from server for download process. Top k value search is used for list the document based on most downloaded file from the consumer.

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ECONOMICAL FEASIBILITY
- TECHNICAL FEASIBILITY
- SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

SYSTEM TESTING AND MAINTENANCE

Testing is vital to the success of the system. System testing makes a logical assumption that if all parts of the system are correct, the goal will be successfully achieved. In the testing process we test the actual system in an organization and gather errors from the new system operates in full efficiency as stated. System testing is the stage of implementation, which is aimed to ensuring that the system works accurately and efficiently.

In the testing process we test the actual system in an organization and gather errors from the new system and take initiatives to correct the same. All the front-end and back-end connectivity are tested to be sure that the new system operates in full efficiency as stated. System testing is the stage of implementation, which is aimed at ensuring that the system works accurately and object.

The main objective of testing is to uncover errors from the system. For the uncovering process we have to give proper input data to the system. So we should have more conscious to give input data. It is important to give correct inputs to efficient testing.Testing is done for each module. After testing all the modules, the modules are integrated and testing of the final system is done with the test data, specially designed to show that the system will operate successfully in all its aspects conditions.

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

UNIT TESTING

Unit testing verification efforts on the smallest unit of software design, module. This is known as "Module Testing". The modules are tested separately. This testing is carried out during programming stage itself. In these testing steps, each module is found to be working satisfactorily as regard to the expected output from the module.

INTEGRATION TESTING

Integration testing is a systematic technique for constructing tests to uncover error associated within the interface. In the project, all the modules are combined and then the entire programmer is tested as a whole. In the integration-testing step, all the error uncovered is corrected for the next testing steps.

V. CONCLUSION

We propose a general expected loss optimization framework for ranking, which is applicable to active learning scenarios for various ranking learners. Under ELO framework, we derive novel algorithms, query level ELO-DCG and document level ELO-DCG, to select most informative examples to minimize the expected DCG loss. We propose a two stage active learning algorithm to select the most effective examples for the most effective queries. We further extend the proposed algorithm to deal with the typical skew grade distribution problem in learning to rank. Extensive experiments on real-world web search data sets have demonstrated great potential and effectiveness of the proposed framework and algorithms. In future, we will investigate how to fuse the query level and document level selection steps in order to produce a more robust query selection strategy. Besides, we will also evaluate our active learning method upon different types of data.

REFERENCES

[1] N. Abe and H. Mamitsuka, "Query learning strategies using boosting and bagging," in Proc. 15th Int. Conf. Mach. Learn., 1998, pp. 1–9.
[2] J. 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.
[3] J. Berger, Statistical Decision Theory and Bayesian Analysis. New York, NY, USA: Springer, 1985.
[4] J. Xu and H. Li, "Adarank: A boosting algorithm for information retrieval," in Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2007, pp. 391–398.
[5] B. Carterette, J. Allan, and R. Sitaraman, "Minimal test collectionsfor retrieval evaluation," in Proc. 29th Annu. Int. ACM SIGIR Conf.Res. Develop. Inform. Retrieval, 2006, pp. 268–275.
[6] W. Chu and Z. Ghahramani, "Extensions of Gaussian processesfor ranking: Semi-supervised and active learning," in Proc. NipsWorkshop Learn. Rank, 2005, pp. 33–38.
[7] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learningwith statistical models," in Proc. Adv. Neural Inf. Process. Syst.,1995, vol. 7, pp. 705–712.
[8] D. Cossock and T. Zhang, "Subset ranking using regression,"in Proc. 19th Annu. Conf. Learn. Theory, 2006, pp. 605–619.
[9] I. Dagan and S. P. Engelson, "Committee-based sampling fortraining probabilistic classifiers," in Proc. 12th Int. Conf. Mach.Learn., 1995, pp. 150–157.
[10] P. Donmez and J. Carbonell, "Active sampling for rank learning via optimizing the area under the ROC curve," in Proc. 31th Eur.Conf. IR Res. Adv. Inform. Retrieval, 2009, pp. 78–89.
[11] P. Donmez and J. G. Carbonell, "Optimizing estimated loss reductionfor active sampling in rank learning," in Proc. 25th Int. Conf.Mach. learn., 2008, pp. 248–255.
[12] V. Fedorov, Theory of Optimal Experiments. New York, NY, USA:Academic, 1972.
[13] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficientboosting algorithm for combining preferences," J. Mach. Learn.Res., vol. 4, pp. 933–969, 2003.
[14] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective samplingusing the query by committee algorithm," Mach. Learn.,vol. 28, nos. 2–3, pp. 133–168, 1997.
[15] J. Friedman, "Greedy function approximation: A gradient boostingmachine," Ann. Statist., vol. 29, pp. 1189–1232, 2001.
[16] T. Fushiki, "Bootstrap prediction and Bayesian prediction undermisspecified models," Bernoulli, vol. 11, no. 4, pp. 747–758, 2005.
[17] D. D. Lewis and W. A. Gale, "A sequential algorithm for trainingtext classifiers," in Proc. 17th Annual Int. ACM SIGIR Conf. ResearchDevelopment Info. Retrieva (SIGIR '9), 1994, pp. 3–12.
[18] D. Lewis and W. Gale, "Training text classifiers by uncertaintysampling," in Proc. 17th Annu. Int. ACM SIGIR Conf. Res. Develop.Inform. Retrieval, 1994, pp. 3–12.
[19] B. Long, O. Chapelle, Y. Zhang, Y. Chang, Z. Zheng, and B. Tseng,"Active Learning for Ranking through Expected Loss Optimization,"in Proc. 33rd Annu.ACMSIGIR Conf., 2010, pp. 267–274.
[20] A. McCallum and K. Nigam, "Employing EM and pool-basedactive learning for text classification," in Proc. 5th Int. Conf. Mach.Learn., 1998, pp. 359–367.

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

[21] B. Settles, "Active learning literature survey," Comput. Sci. Dept.,Univ. of Wisconsin, Madison, WI, USA, Tech. Rep. 1648, 2009.
[22] F. Xia, T.-Y. Liu, J. Wang, W. Zhang, and H. Li, "Listwiseapproach to learning to rank: theory and algorithm," in Proc. 25thInt. Conf. Mach. Learn., 2008, pp. 1192–1199.
[23] L. Yang, L. Wang, B. Geng, and X.-S. Hua, "Query sampling forranking learning in web search," in Proc. 32nd Int. ACM SIGIRConf. Res. Develop. Inform. Retrieval, 2009,
pp. 754–755.
[24] E. Yilmaz and S. Robertson, "Deep versus shallow judgments inlearning to rank," in Proc. 32nd Int. ACM SIGIR Conf. Res. Develop.Inform. Retrieval, 2009, pp. 662–663.
[25] H. Yu, "SVM selective sampling for ranking with application todata retrieval," in Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery.Data Mining, 2005, pp. 354–363.