ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

IARJSET

International Advanced Research Journal in Science, Engineering and Technology

National Conference on Innovative Applications and Research in Computer Science and Engineering (NCIARCSE-2017)

AGTI's Dr. Daulatrao Aher College Engineering, Vidyanagar Extension, Karad

Vol. 4, Special Issue 4, January 2017

Modified Active Learning for Document Level Clustering

Snehal Patil[1], Prof. Jayant Jadhav[2]

Student, SAOE, Pune, India [1]

Assistant Professor, SAOE, Pune, 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 a two-stage ELO-DCG algorithm which incorporate both query and document selection into active learning.

Keywords: Active learning, ranking, expected loss optimization.

1. 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 time consuming to label the samples. Existing algorithms for learning to rank may be categorized into three groups: point wise approach [8], pair wise approach [26], and listwise approach [22]

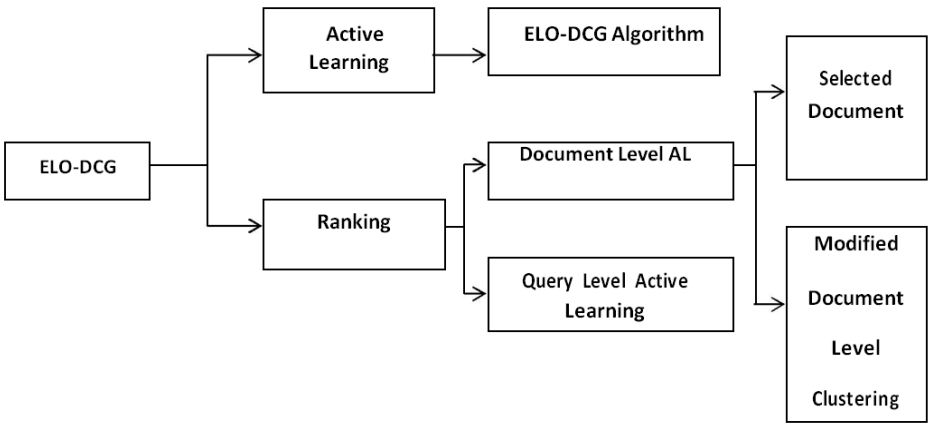

Fig. Proposed Architecture

2. MOTIVATION

The main motivation for active learning is that it usually requires time and/or money for the human expert to label examples and those resources should not be wasted to label non-informative samples, but be spent on interesting ones.

Optimal experimental design [12] is closely related to active learning as it attempts to find a set of points such that the variance of the estimate is minimized. In contrast to this "batch" formulation, the term active learning often refers to an incremental strategy [7].

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
National Conference on Innovative Applications and Research in Computer Science and Engineering (NCIARCSE-2017)
AGTI's Dr. Daulatrao Aher College Engineering, Vidyanagar Extension, Karad
Vol. 4, Special Issue 4, January 2017

3. SCOPE OF SYSTEM

The system can be used for the searching the documents based on the proposed work to achieve the better performance in document level clustering

4. ACTIVE LEARNING FOR RANKING THROUGH EXPECTED LOSS OPTIMIZATION

Active learning for ranking through expected loss optimization select most informative examples for ranking learning. In ranking represents an important class of supervised ma-chine learning tasks with the goal of automatically constructing ranking functions from training data [1].

Expected loss optimization As explained in the previous section, a natural strategy for active learning is based on variance minimization. The Variance, in the context of regression, stems from the uncertainty in the prediction due to the finiteness of the training set.

Cohn et al. [7] Proposes to select the next instance to be labeled as the one with the highest variance. In the case of ranking, the input instance is a query and a set of documents associated with it, while the output is a vector of relevance scores.

If the query q has n documents, let us denote by Xq=(x1 …xn) the feature vectors describing these (query, document) pairs and by Y = (y1…yn) their labels. As before we have a predictive distribution P (Y|Xq, D,). Unlike active learning for classification and regression, active learning for ranking can select examples at different levels. One is query level, which selects a query with all associated documents Xq; the other one is document level, which selects documents xi individually.

1 Query Level
In the case of ranking, the ‖action‖ in ELO framework is slightly different than before because we are not directly

Interested in predicting the scores, but instead we want to produce a ranking. So the set of actions is the set of permutations of length n and for a given permutation p, the rank of the ith document. The expected loss for a given p can thus be written as: ∫

Equation 1
The next section will detail the computation of the expected loss where is the DCG loss. As before, the ELO principle for active learning tells us to select the queries with the highest expected losses: ∫

Equation 2
As an aside, note that the ranking minimizing the loss (3) is not necessarily the one obtained by sorting the documents according to their mean predicted scores. This has already been noted for instance in [27].

2.2 Document Level
Selecting the most informative document is a bit more complex because the loss function in ranking is defined at the query level and not at the document level.

We propose a new approach for document level active learning. In our proposed method **Documents** in a collection are assigned **terms** from a set of n terms. The **term vector space** W is defined as:

if term k does not occur in document di, wik
= 0

if term k occurs in document di, wik is greater than zero (wik is called the **weight** of term k in document di)

Similarity between di and dj is defined as:

$$\sum$$

Equation 3
Where **d**i and **d**j are the corresponding weighted term vectors and |**d**i| is the length of the document vector **d**i.

In our proposed approach we extend the Similarity measure for obtaining an expected loss. We calculate the expected loss incurred during computation of similarity between document di and dj.

$$EL(di, dj) = 1 - (\cos(di, dj))$$

We consider the less value of loss during computation for the better document matching.

Feature of algorithm:
1) Reduce labeling effort.
2) Required less time and money

Problem in algorithm:
1) It perform ranking only on query not on document.
2) It mixes two type of uncertainties, the one stemming from the noise and variance.

5. OPTIMIZING SEARCH ENGINES USING CLICK THROUGH DATA

Optimizing search engine presents an approach to learning retrieval functions by analyzing which links the users click on in the presented ranking. In this define what click through data is, how it can be recorded and how it can be used to generate training examples in the form of preferences? Click through data can provide training data in the form of relative preferences. Based on a new formulation of the learning problem in information retrieval, this derives an algorithm for learning a ranking function [7]

Feature of algorithm:
1) Click through data easily recorded and it required less cost.
2) Each query assigns unique ID in query log with query word in presented ranking.

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
National Conference on Innovative Applications and Research in Computer Science and Engineering (NCIARCSE-2017)
AGTI's Dr. Daulatrao Aher College Engineering, Vidyanagar Extension, Karad
Vol. 4, Special Issue 4, January 2017

Problem in algorithm:
1) No consideration for user personal preferences.

Feature of algorithm:
1) Reduce labeling effort.
2) Required less time and money

Problem in algorithm:
3) It perform ranking only on query not on document.
4) It mixes two types of uncertainties, the one stemming from the noise and variance.

Feature of algorithm:
3) Click through data easily recorded and it required less cost.
4) Each query assign unique ID in query log with query word in presented ranking.

Problem in algorithm:
1) No consideration for user personal preferences.
2) Ranking algorithms mainly based on similarity. Similarity between pages (Page Rank).

6. ADARANK: A BOOSTING ALGORITHM FOR INFORMATION RETRIEVAL

Adarank develop a new learning algorithm that can directly optimize any performance measure used in document retrieval. In document retrieval, usually ranking results are evaluated in terms of performance measures such as MAP (Mean Average Precision) and NDCG (Normalized Discounted Cumulative Gain).

AdaRank algorithm can iteratively optimize an ex-ponential loss function based on any of IR performance measures. AdaRank can be viewed as a machine learning method for ranking model tuning [5].

Feature of algorithm:
1) High accuracy in ranking.
2) Easy in implementation.

Problem in algorithm:
1) It does not improve ranking accuracy in performance measurements.
2) It is not time consuming.

7. SVM SELECTIVE SAMPLING FOR RANKING WITH APPLICATION TO DATA RETRIEVAL

It produced practical applications in information retrieval. SVM selective sampling technique is used for learning ranking function. Selective sampling technique is to select the most ambiguous samples for ranking at each round, so that the users feedback on those samples will maximize the degree of learning. In this sampling technique significantly reduces the labeling effort to learn an accurate SVM ranking function and it apply method to data retrieval application [2]

Feature of algorithm:
1) SVM sampling reducing the labeling effort of informative sample.
2) SVM achieve high accuracy.

Problem in algorithm:
1) Data pairs which are closed is more a ambiguous.
2) For classification acquiring large number of training (labeled) data is expensive or hard.
3) For classification acquiring large number of training (labeled) data is expensive or hard.

Previous System Working Graph:

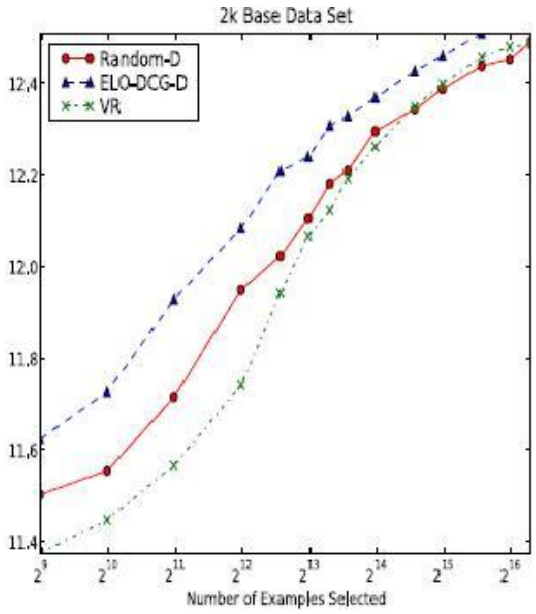

Fig.1

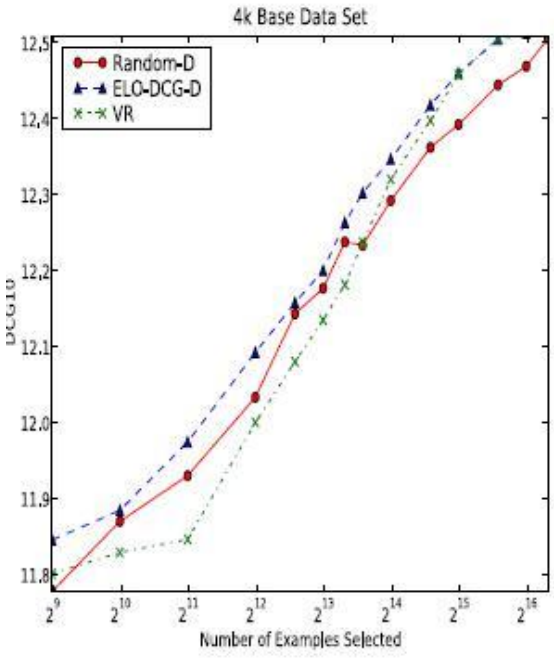

Fig.2

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
National Conference on Innovative Applications and Research in Computer Science and Engineering (NCIARCSE-2017)
AGTI's Dr. Daulatrao Aher College Engineering, Vidyanagar Extension, Karad
Vol. 4, Special Issue 4, January 2017

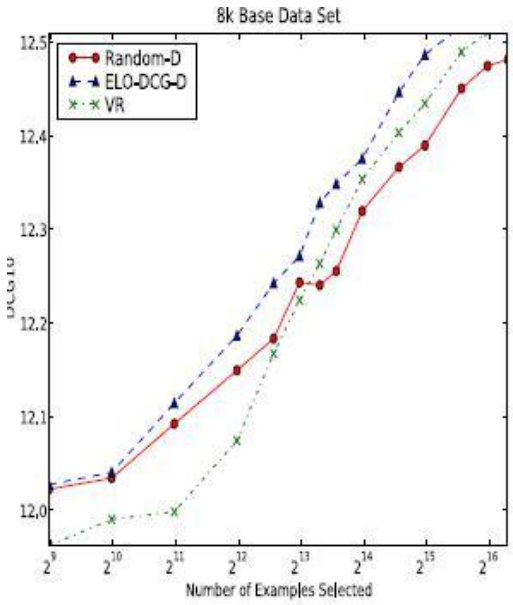

Fig.3

Figures 1,2,3 shows the DCG comparison of document level ELO-DCG, variance reduction based document selection, and random document selection with base sets of sizes 2,4, and 8k shows that ELO-DCG algorithm outperforms the other two document selection methods at various sizes of selected examples.

PROPOSED RESULT

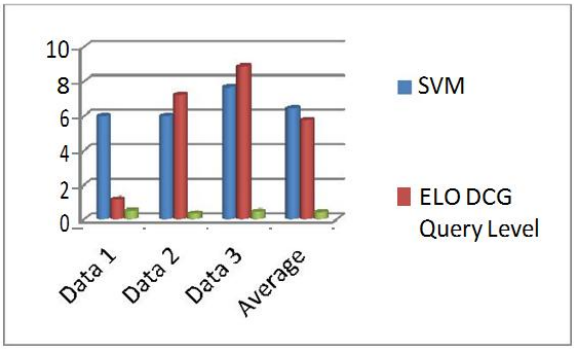

Fig.4 Proposed Graph for loss

8. CONCLUSION

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] C. Campbell, N. Cristianini, and A. Smola, "Query learning with large margin classifiers," in Proc. 17th Int. Conf. Mach. Learn., 2000, pp. 111–118.
[5] B. Carterette, J. Allan, and R. Sitaraman, "Minimal test collections for 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 processes for ranking: Semi-supervised and active learning," in Proc. Nips Workshop Learn. Rank, 2005, pp. 33– 38.
[7] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with 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 for training 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 reduction for 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 efficient boosting 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 sampling using the query by committee algorithm," Mach. Learn., vol. 28, nos. 2–3, pp. 133–168, 1997. [15] J. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Statist., vol. 29, pp. 1189– 1232, 2001.
[15] T. Fushiki, "Bootstrap prediction and Bayesian prediction under misspecified models," Bernoulli, vol. 11, no. 4, pp. 747– 758, 2005.
[16] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in Proc. 17th Annual Int. ACM SIGIR Conf. Research Development Info. Retrieva (SIGIR '9), 1994, pp. 3–12.
[17] D. Lewis and W. Gale, "Training text classifiers by uncertainty sampling," 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.
[18] A. McCallum and K. Nigam, "Employing EM and pool-based active learning for text classification," in Proc. 5th Int. Conf. Mach. Learn., 1998, pp. 359–367.
[19] B. Settles, "Active learning literature survey," Comput. Sci. Dept., Univ. ofWisconsin, Madison, WI, USA, Tech. Rep.1648, 2009.
[20] 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– 1199.
[21] L. Yang, L. Wang, B. Geng, and X.-S. Hua, "Query sampling for ranking learning in web search," in Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2009, pp. 754–755.
[22] E. Yilmaz and S. Robertson, "Deep versus shallow judgments in learning to rank," in Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2009, pp. 662–663.
[23] H. Yu, "SVM selective sampling for ranking with application to data retrieval," in Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery. Data Mining, 2005, pp. 354–363.
[24] Z. Zheng, H. Zha, T. Zhang, O. Chapelle, K. Chen, and G. Sun, "A general boosting method and its application to learning ranking Functions for web search," in Proc. Adv. Neural Inf. Process. Syst. 20, 2008, pp. 1697–1704.
[25] O. Zoeter, N. Craswell, M. Taylor, J. Guiver, and E. Snelson, "A decision theoretic framework for implicit relevance feedback," in Proc. NIPS Workshop Mach. Learn. Web Search, 2007, pp. 133–142.
[26] T. Joachims, "Optimizing search engines using clickthrough data," in

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

National Conference on Innovative Applications and Research in Computer Science and Engineering (NCIARCSE-2017)

AGTI's Dr. Daulatrao Aher College Engineering, Vidyanagar Extension, Karad

Vol. 4, Special Issue 4, January 2017

Proc. 8th ACM SIGKDD Int. Conf Knowl. Discovery Data Mining, 2002, pp. 133–142.

[27] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in Proc. 22nd Int. Conf. Mach. Learn. 2005, pp. 89–96.

[28] Y. Freund, R. Iyer, R. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," in Proc. 15th Int. Conf. Mach. Learn., 1998, pp. 391–398.

[29] 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.

[30] C. Cortes, M. Mohri, and A. Rastogi, "Magnitude-preserving ranking algorithms," in Proc. 24th Int. Conf. Mach. Learn., 2007, pp. 169–176.

[31] Z. Zheng, K. Chen, G. Sun, and H. Zha, "A regression framework for learning ranking functions using relative relevance judgments," in Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2007, pp. 287–294.

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

[33] S. Rodrigo, M. A. Gonalves, and A. Veloso, "Rule-based active sampling for learning to rank," in Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases, 2011, pp. 240–255.

[34] 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.tive Learning for Ranking through Expected Loss Optimization Bo Long, Jiang Bian, Olivier Chapelle, Ya Zhang, Yoshiyuki Inagaki, and Yi Chang. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 5, MAY 2015