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 12, December 2016

Modified Active Learning for Document Level Clustering

Vaishali Balu Patole [1], Suvarna Potdukhe [2]

P.G. Student, Department of Information Technology, RMD Sinhgad School of Engineering, Warje, Pune, India[1]

Assistant Professor, Department of Information Technology, RMD Sinhgad School of Engineering, Warje, Pune, India[2]

ABSTRACT*:* Active learning for ranking is a new trend used in web search ranking, recommendation system and an online advertising. But in ranking the performance of the ranking is strongly affected by the number of labelled examples. But in case of training obtaining the trained dataset is a time consuming and expensive task. Existing system is proposed using an expected loss optimization. Different challenges are discussed in this paper. The proposed system is an implemented using active learning using an unsupervised approach. Here our approach will be able learn from automatically from the data samples and finds an expected loss using proposed function. The proposed approach will be able to find the similar document using proposed cosine similarity measure.

KEYWORDS*:* Active learning, An Expected Loss Optimization, Ranking.

I. INTRODUCTION

In many information retrieval problem ranking is a major issue that needs to be addressed. Ranking plays an important role in web search, advertising and recommendation system. Learning to rank presents supervised machine learning task with automatically constructing ranking functions from training data. Amount of labelled data represents the quality of ranking in supervised machine learning problems. To achieve high quality ranking in supervised ranking problems are the large amount training data is required. As it is known that the labelling of data samples is time consuming and an expensive task. Active learning is new technique used to avoid the labelling efforts in supervised learning. The active learning is mostly studied in the field of classification. Active learning algorithms are studied into following groups as-

1. Point wise approach
2. Pair wise approach and
3. List wise approach.

Some active learning approach such as query-by-committee (QBC) has not been justified for the ranking tasks under regression framework [1]. In supervised learning each data sample can be treated completely independently.

In ranking problems are data items conditionally dependant on other data items. In case of document retrieval number of irrelevant document is of orders of magnitude more than that of relevant document. It is important to consider the skewness while selection of data for ranking, therefore there is need of active learning.

The system proposes a general active learning framework called an expected loss optimization (ELO) and apply it to ranking. The proposed framework gives a loss function and the samples minimizing the expected loss (EL) are the most informative ones [1]. This framework derives a novel active learning algorithm for ranking. An algorithm uses a function to select most informative examples that reduces a chosen loss [1].

This paper uses web search ranking problem as an example to demonstrate the ideas and perform evaluation. The proposed method is generic and may be applicable to all ranking applications [1]. In web search ranking system minimized an expected discounted cumulative gain (DCG) loss. The proposed algorithm may be easily adapted to other ranking such as normalized discounted cumulative gain (NDCG) and average precision. To solve the problem of the data dependency proposed algorithm is further an extended to a two stage active learning. This algorithm seamlessly integrates query level and document level data selection [1]. At last system an extended a algorithms to solve the skew grade distribution problem [1].

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 12, December 2016

II. RELATED WORK

The simplest and most common strategy is uncertainty sampling, where the active learning algorithm queries points for which the label uncertainty is highest. The drawback of this type of approach is that it mixes two types on uncertainties, the one stemming from the noise and the variance. The noise is something occurs to learning problem which does not depend on the size of the training set.
An active learner should not spend too much effort in querying points in noisy regions of the input space uncertainty in the model parameters resulting from finiteness of the training set. Active learning should thus try to minimize this variance and this was first proposed in [3]. These learning to rank approaches are capable of combining different kind of features to train ranking functions. The problem of ranking can be formulated as that of learning a ranking function from pair-wise preference data. The idea is to minimize the number of contradicting pairs in training data. For example, Rank SVM [4] uses support vector machines to learn a ranking function from preference data. Rank Net [5] applies neural network and gradient decent to obtain a ranking function. Rank Boost applies the idea of boosting to construct an efficient ranking function from asset of weak ranking function.
The studies reported in[6] proposed framework call GB Rank using gradient decent in function spaces, which is able to learn relevant ranking information in the context of web search. The main focus of this paper, active learning for ranking, compared with traditional active learning, there is still limited work in literature. Donmez and Carbonell studied the problem of document selection in ranking [7]. Their algorithm selects the documents which, once added to the training set, are the most likely to result in a large changes in the model parameters of the ranking function. They can apply their algorithm to Rank SVM and Rank Boost also in the context of Rank SVM [2] suggests to add the most an ambiguous pairs of documents to the training set, that is documents whose predicted relevance scores are very close under the current model. In case of binary relevance, proposed a greedy algorithm which select document that are the most likely to differentiate two ranking system in terms of an average precision**.**
There are some related works about query sampling. Yilmaz and Robertson [9] empirically show that having more queries but less number of documents per query is better than having more documents and less queries. Yang long et al. active learning for ranking through an expected loss optimization et al. proposes a greedy query selection algorithm that tries to minimize a liner combination of query difficulty, query density and query diversity [8].

III. APPROACHE USE

A. An Expected Loss Optimization for Active Learning

The natural strategy for active learning is based on the variance minimization. Due to finite training data set it leads into the uncertainty in prediction. The reference [7] proposes to select the next instance to be labelled as the one with the highest variance. This approach applied only regression; instead we apply Bayesian an expected loss to generalize this. There are some challenges occurs as-

- There is no notion of classification in ranking; therefore margin based active learning algorithms are not readily applicable.
- In most of the supervised learning approach, each data samples are treated independently.
- Ranking problems are often associated with much skewed data distributions.

We proposed active learning framework for an expected loss optimization and apply it for the ranking. The proposed algorithm further an extended to two stage active learning to integrate a query level and document level data selection. We propose a modified approach for document level clustering and ranking. Our approach works for the document being selected by the user during its single session.

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 12, December 2016

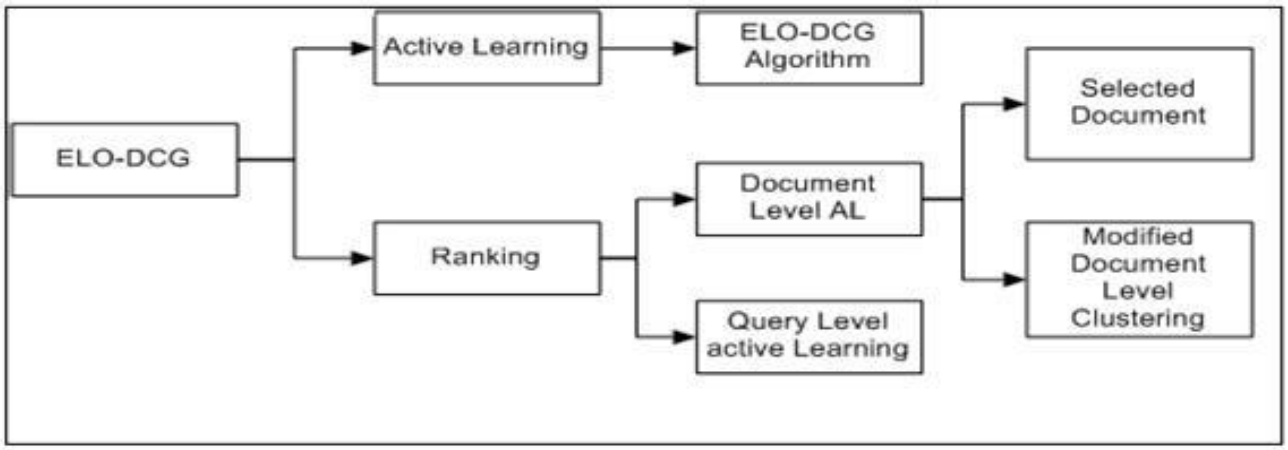

Fig1- Proposed Approach of Active Learning

The proposed work is in fig1. Presents the general active learning framework based on expected loss optimization. System derives a novel algorithm to select most an informative example by optimizing the expected discounted cumulative gain loss. System proposed a two stage active learning algorithm for ranking, the motivation behind the active learning is that, it usually requires more time and money to label the examples. This wastage of time must be avoided or used for the labelling of informative examples. Optimal experimental design is closely related to active learning as it attempts to find a set of points such that the variance of an estimate is minimized [1].

IV. PROPOSED METHODOLOGY

A. *DOCUMENT LEVEL ACTIVE LEARNING*

In active learning selection of most informative document is a bit complex task, because loss function in ranking is defined at the query level and not at the document level. In this work we proposed a new approach for document level clustering using an active learning. 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:

$$\cos(\mathbf{d}_i, \mathbf{d}_j) = \frac{\sum_{k=1}^{n} w_{ik} w_{jk}}{|\mathbf{d}_i| \; |\mathbf{d}_j|}$$

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 this proposed approach we extend the similarity measure for obtaining an expected loss. We calculate an 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.

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 12, December 2016

B. ***PROPOSED ALGORITHM***

Input : Document set D={d1,d2,….,dn}, Query Q,

Output : Selected document set D={d1,d2,…,dk}

Method :

Search query q on search engine.
Obtain the search result document and store in DB.
For document j=1to n in DB
Obtain similarity measure cos(di,dj);
Obtain the expected loss EL(di,dj);
Store EL for all dj in DB
End for
For all EL values in DB
Sort all values in descending order.
End for
Select top values from EL which minimizes the loss.
Select document from top EL values.
Return top k documents.

The algorithm is working to fuse the both query and document to get more accurate query document results for user selected query.

V. EXPERIMENTAL RESULT

The experimental result can shows the graph of the ranking of user click query and user clicked sorted document which is drown on the basis of considering the user click query on search engine. The graph and the values of graph are shown below,

Table 1: User Clicked Query Ranking

Search Keyword Name	Search Query on	Query List	User click Query	Rank Query by User Clicked
Hiranandani	Google	50	10	0
Hiranandani	Search Engine	50	10	10

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 12, December 2016

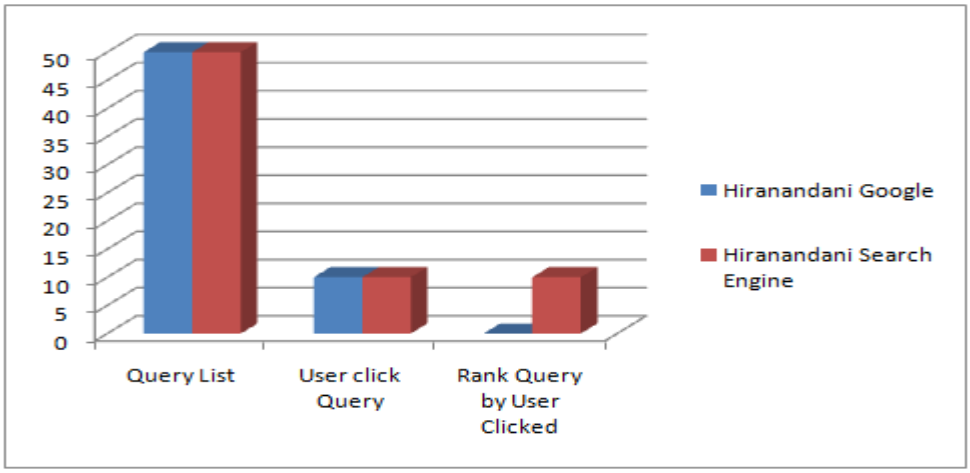

Fig- Graph of Ranking only User Clicked Query

The above graph shows the user clicked query ranking, in this user searching a keyword Hiranandani on both i.e. on Google and on Search engine. It shows the fifty query document results for Hiranandani query. In this search user click on ten queries as per the requirements, so it can only rank ten queries of Hiranandani only on search engine, it does not rank the Hiranandani queries on Google and it rank zero query for Google.

Table 2 : Sorted Document Results of only User Clicked Query

Search Keyword	User Clicked Query	Sorted Query Document in Descending Order
Vodafone	10	50
Tata	15	50
Reliance	20	50

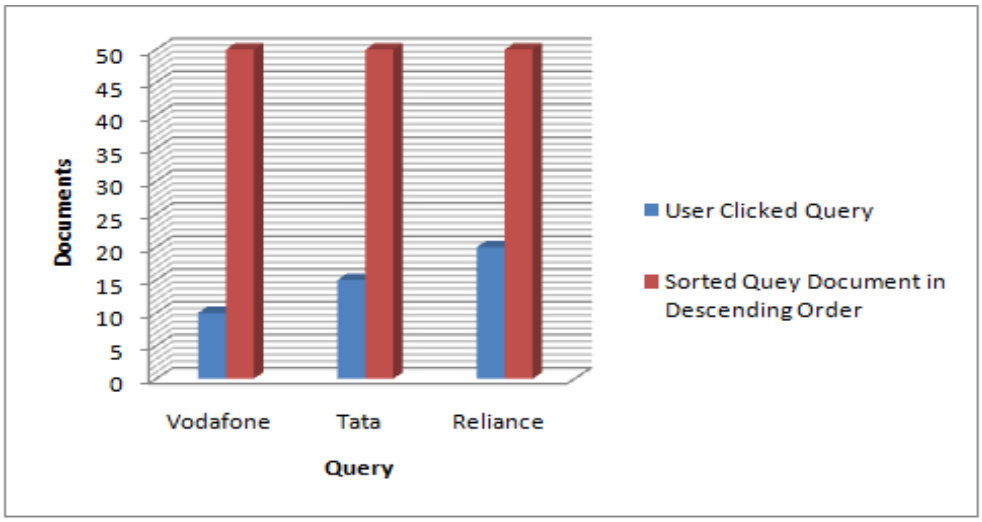

Fig- Sorted Document Results of Clicked Query in Descending Order

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 12, December 2016

The above graph shows the sorted document results of clicked query in descending order. In this user searching a keyword Vodafone, Tata, Reliance on both i.e. on Google and on Search engine. It shows the fifty query document results for each keyword. In this for Vodafone keyword user click on ten query it gives all fifty sorted document results in descending order for that ten query and for Tata keyword user click twenty query so, it also give fifty sorted document results in descending order.

VI. CONCLUSION

The proposed system is implemented using an expected loss optimization based on active learning. The system is proposed for the document level clustering using proposed cosine similarity based measure. The system uses the proposed approach for finding an expected loss minimization function.

REFERENCES

[1] Bo Long, Jiang Bian, Olivier Chapelle, Ya Zhang, Yoshiyuki Inagaki, and Yi Chang , "Active Learning for Ranking through Expected Loss Optimization", IEEE transactions on knowledge and data engineering, vol. 27, no. 5, may 2015.
[2] H. Yu, "SVM selective sampling for ranking with application to data retrieval", in Proc. 11th ACM SIGKDD Int. Conf. Knowledge. Discovery. Data Mining, 2005, pp.354-363.
[3] 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.
[4] T. Joachims, "Optimizing search engines using click through data", in Proc. 8th ACM SIGKDD Int. Conf. Knowledge. Discovery Data Mining, 2002, pp. 133-142.
[5] 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.
[6] 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.
[7] 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.
[8] L. Yang, L. Wang, B. Geng, and X.-S. Hua, "Query sampling for ranking learning in web search", in Proc. 32nd Int. ACM SIGIRConf. Res. Develop. Inform. Retrieval, 2009, pp. 754–755.
[9] 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.