Volume 3, Issue 8, August 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Global Ranking Model for Domain Dependent Information Retrievals

P.M.R.Rajeswari , V.Ganesh Datt, T.Sasi Krishna

Department of CSE & JNTUK

India

Abstract- Ranking allows users to have best results. As search engines return huge data, surfing that for required information is time consuming. Designing and implementing a ranking model for every vertical domain is not a feasible solution. In this paper we propose a mechanism that allows an algorithm to adapt to new domains automatically. This will help in using the same mechanism to every new domain. This adaptation provides a good solution or domain specific search. This kind of algorithm can reduce the development time and cost. We also proposed a measurement for adaptation and built a prototype. The experimental results reveal that the proposed mechanism makes the algorithm to adapt to various domains.

Index Terms – Ranking model, ranking adaptation, information retrieval, domain specific search

I. INTRODUCTION

Day by day huge amount of data is being added to Internet. This is of course helping to fulfill the information needs of users. The information retrieval systems like search engines, due to the availability of data, return huge amount of data. Users are supposed to browse such data to get required information. This causes the wastage of user time. To overcome this drawback ranking techniques are used to provide precise information that can be used by users directly without spending lot of time surfing search results. Ranking models can reduce the time taken to know the satisfactory information. Many ranking models depend on learning models. There are many such learning models namely ListNet [2], RankBoost [4], LambdaRank [1], SVM [5], [6], and RankNet [3]. These are proven solutions for ranking models in the context of searching through crawlers of WWW (World Wide Web). In case of search engines, some of them are domain specific in nature. They have been built to shift from broad based search to domain specific search. They are known as vertical search engines that focus on specific information retrieval. For instance there are search engines for medical data, chemical data, bioinformatics, science, education etc. This paved the way for domain specific search and users can get specialized information. Domain specific search engines operate on different documents and different formats which are compatible in that domain. Text search techniques are generally used by broad based search engines. They also treat images as text based so as to retrieve them easily. Term Frequency is also used by them for ranking available documents. The broad based models for ranking work for multiple domains in general. However, they are not specialized for domain specific search. They do not take search word as it is. Instead, they can use semantic search that makes it to consider words with similar meaning as well. In fact they use lexical analysis to retrieve results. Building a model for every domain is also not a viable solution as it is costly and time consuming. Users are to switch from one model to another model. Many experiments have been found in the literature on broad based search. This kind of search gives information that can be used by web users. The problem with such model is that it becomes so generic and can't be used for domain specific search. For domain specific search, it is best to use a ranking adaptation model that allows users to have domain specific search. Such model which can adapt to different domains can make it a viable solution. However developing such model which is flexible and domain specific when adapted is a challenging task.

Classifier adaptation can be compared with adaptation of ranking for new domains. The research in this area suggest some comparisons [7], [8], [9], [10], [11]. However, the ranking model which can adapt to different domains is relatively new concept. There was no solution found in the literature as it is a new thing in the research. Instead, the classifier adaptation and other technique such as concept drifting [12] were available in the literature. The former deals with a corpus of documents for adapting the classifier to different documents. The latter is used to handle concept drifting in the information retrieval systems. In this paper, we use labeled data for making a ranking model which can be used for different domains as it can adapt automatically. This feature has got significance as it is flexible and time saving solution when compared with other alternative such as classifier adaptation. As this model uses the labels of data, it can adapt to new domains so as to provide users the provision for domain specific search without the need for having many search mechanisms. The remainder of the paper is structured into sections. Section 2 focuses on review of literature. Section 3 provides information about the new ranking adaptation model. Section 4 gives details about the experimental results while section 5 provides conclusions.

II. Related Work

Many researches came into existence in the literature for ranking search results. In this section we present the ranking models are very close to the technique used in this paper. The models in the literature are broadbased techniques. They include classical BM 25 [15], and information retrieval [13], [14]. These models need some parameters to be adjusted. There are many different learning methods available. They use labeled data and they have complicated features.

This is the motivation behind the work that needs to develop a model that adapts a model with new models. Towards this many algorithms came into existence of late. These techniques learn to rank results and present data in a useful manger. Users can use such data easily without wasting their time.

Some techniques also transform the ranking problems into classification problems. They work on pairs of documents for labeling. Many such algorithms are found in the literature. They include SVM [5], [6], ListNet [2], BankBoost [4], RankNet [3] and LambdaRank [1]. All these techniques focused on optimization of ranking. However in this paper we build a new model for ranking. We actually build a ranking model which can adapt to new domains. Instead of developing a new model for every search domain, also known as vertical search domain, the proposed algorithm can adapt to new domains with ease. This adaptation makes it very useful. It also saves lot of development time and cost. A mixture model later was presented by DAume and Marco. This model addresses the problems associated with classifiers. They also find the differences between training and test sets [16]. For this technique a model is built [8] for boosting. Blitzer et al. also presented a structural correspondence learning method [7]. Different domains can be mined using this. Yang et a. [10] developed a ranking model that ranks cross domain videos. This algorithm is known as adaptive SVM. For designing classification problems all these algorithms can be used. This paper focuses on developing a ranking model which can be effectively adapted to various domains seamlessly. This model has been tested. The emperical results revealed that the model is working fine.

III. Raking Adaptation

This section gives details about the ranking model which can adapt to new domains. First, adaptation problem is described here. A set of queries is denoted by Q, D denotes a set of documents. Human annotators are used to annotate search results. There are other ranking models such as PageRank [18] and HITS [17]. Learning to rank is the theme used in the proposed algorithm. This is an important phenomenon as learning to rank can made different and also allows us to adapt the model to new domains. In the implementation, we assume the return docs and the documents in training set are reasonably small. Labeled data set is used by intermediate ranking models in order to adapt to new domains.

Ranking Adaptation SVM

We also assume that the new domains are smaller in size. Therefore it is possible for the algorithm to adapt to new models with ease. With required assumptions in mind it is possible to use datasets that need prior knowledge for ranking model adaption. The traditional frameworks such as neural networks [20] and SVM [19] has certain problems in handling domain specific data in the adaptation model. This problem can be solved by using a regularization framework that can adapt to new ranking models. The proposed model is as shown below.

$$\min = \frac{1-\partial}{2}\|f\|^2+\frac{\partial}{2}\sum_{r=1}^{R} \Theta r\|f-f^a_r\|^2+C\,\Sigma\epsilon_{ijk}$$

$$\text{s.t. } f(\varpi(q_i, d_{ij})) - f(\varpi(q_i, d_{ij})) > 1-\epsilon_{ijk} \quad \epsilon_{ijk} > 0, \text{ for}$$

$$\forall i \in \{1,2,\ldots\ldots,M\}, \forall j\forall k \in \{1,2,\ldots,n(qi)\} \text{ with } yij > yik$$

Adapting to Multiple Domains

The algorithm which has been proposed can be extended to support many domains with ease. This will facilitate the new domains to be used for searching. The domain specific vertical search is possible with the new model. The adaptation to multiple domains can be deduced as follows.

$$\min = \frac{1-\partial}{2}\|f\|^2+\frac{\partial}{2}\sum_{r=1}^{R} \Theta r\|f-f^a_r\|^2+C\,\Sigma\epsilon_{ijk}$$

$$\text{s.t. } f(\varpi(q_i, d_{ij})) - f(\varpi(q_i, d_{ij})) > 1-\epsilon_{ijk} \quad \epsilon_{ijk} > 0, \text{ for}$$

$$\forall i \in \{1,2,\ldots\ldots,M\}, \forall j\forall k \in \{1,2,\ldots,n(qi)\} \text{ with } yij > yik$$

The data comes from domains are generally having different features. The model which has been proposed in this paper can make use of those features. This will make it suitable for adapting new models easily. In the process, the ranking loss concept also made it more consistent. The result is based on the comparison of similar documents. Slack scaling is also used to make the technique robust. The other aspects considered are optimization problem and also margin rescaling. The equation for this sis as given below.

$$\min = \frac{1-\partial}{2}\|f\|^2+\frac{\partial}{2}\sum_{r=1}^{R} \Theta r\|f-f^a_r\|^2+C\,\Sigma\epsilon_{ijk}$$

$$\text{s.t. } f(\varpi(q_i, d_{ij})) - f(\varpi(q_i, d_{ij})) > 1-\epsilon_{ijk}-\sigma_{ijk} \quad \epsilon_{ijk} > 0, \text{ for}$$

$$\forall i \in \{1,2,\ldots\ldots,M\}, \forall j\forall k \in \{1,2,\ldots,n(qi)\} \text{ with } yij > yik$$

Slack rescaling formula is as given below.

$$\text{Max} - 1/2 \sum_{ijkk}^{n} \quad \sum_{ijkk}^{n} \quad \alpha ijk\ \alpha lmn\ X^{T}_{ijk} X_{lmn}$$

$$+\sum_{ijkk}^{n} \quad (1-\delta f^{2}(X_{ijk}))\ \alpha ijk$$

$$\text{s.t. } f(\varpi(q_i, d_{ij})) - f(\varpi(q_i, d_{ij})) > 1 - \epsilon_{ijk} - \sigma_{ijk}\ \epsilon_{ijk} > 0, \text{ for}$$

$$\forall i \in \{1,2,\ldots\ldots,M\}, \forall j \forall k \in \{1,2,\ldots,n(qi)\} \text{ with } yij > yik$$

IV. Experimental Results

In this paper, the experiments are carried out using a prototype application. The application is built using technologies like Servlets, JSP, and JDBC etc. The environment used is a PC with Core 2 Duo processor, 4GB of RAM. The datasets are TD 2004 and TD 2003. The proposed model is tested with these datasets. The performance is evaluated using measures known as mean average precision and normalized discounted cumulative gain. The result revealed that the proposed ranking adaptation model is working fine and can be used in real world applications.

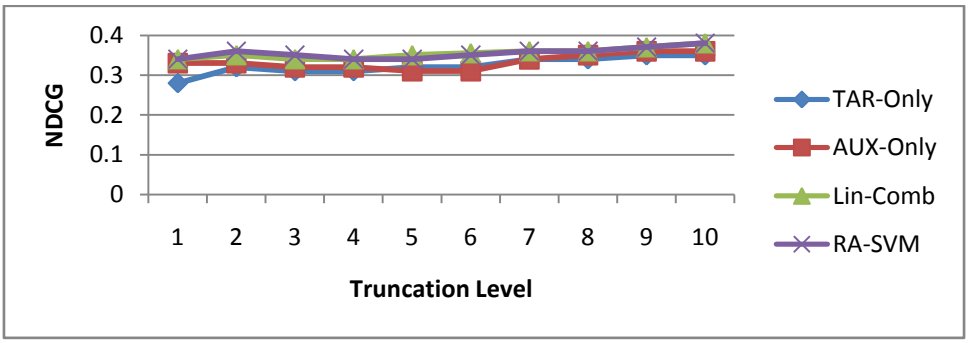

Fig. 1 - TD2003 to TD2004 adaptationwith five queries

As seen in fig. 1, adaptation performance is compared. The two dataset's results are compared with five queries. The proposed system outperforms the existing systems.

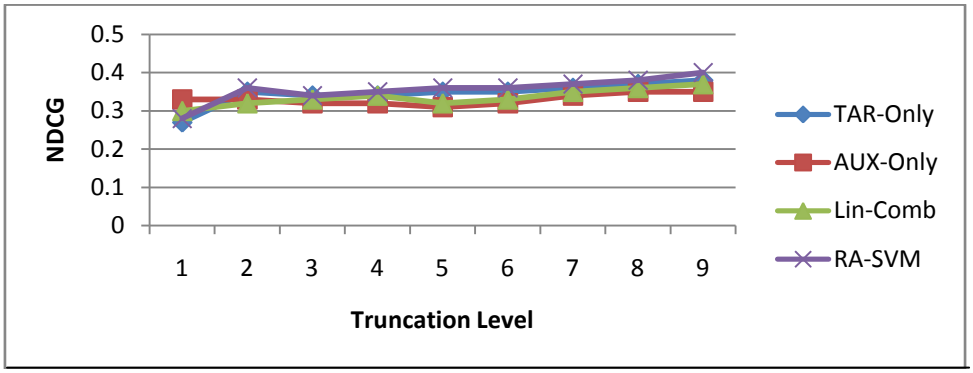

Fig. 2 - TD2003 to TD2004 adaptation with ten queries

As seen in fig. 2, adaptation performance is compared. The two dataset's results are compared with five queries. The proposed system outperforms the existing systems.

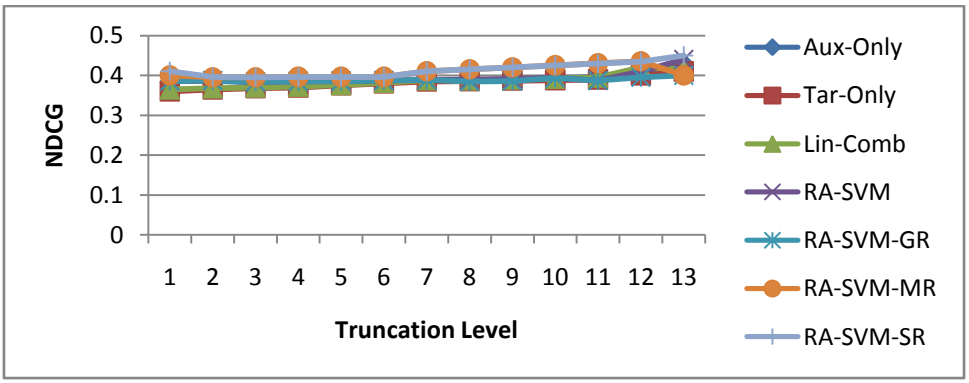

Fig. 3 –NDCG Results of web page search to image search adaptation with five labeled queries

As seen in fig. 3, adaptation performance is compared. The two dataset's results are compared with five queries. The proposed system outperforms the existing systems.

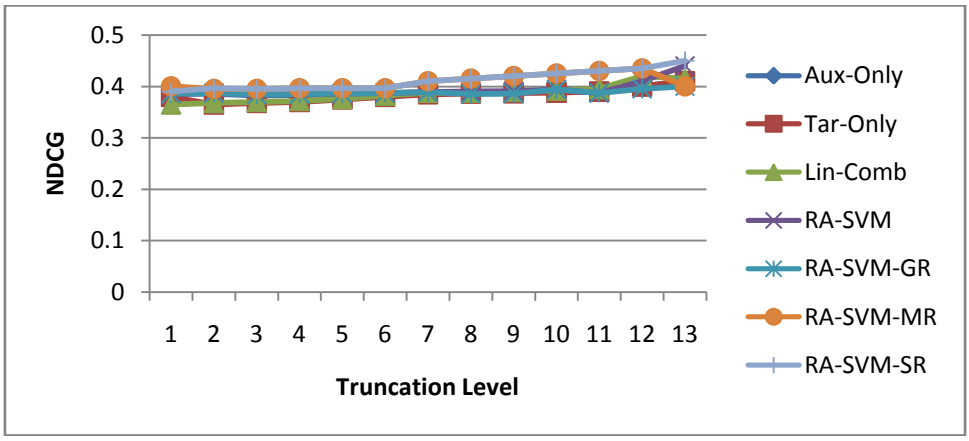

Fig. 4 – NDCG Results of web page search to image search adaptation with ten labeled queries

As seen in fig. 4, adaptation performance is compared. The comparison is truncation level vs. NDCG. The proposed system outperforms the existing systems.

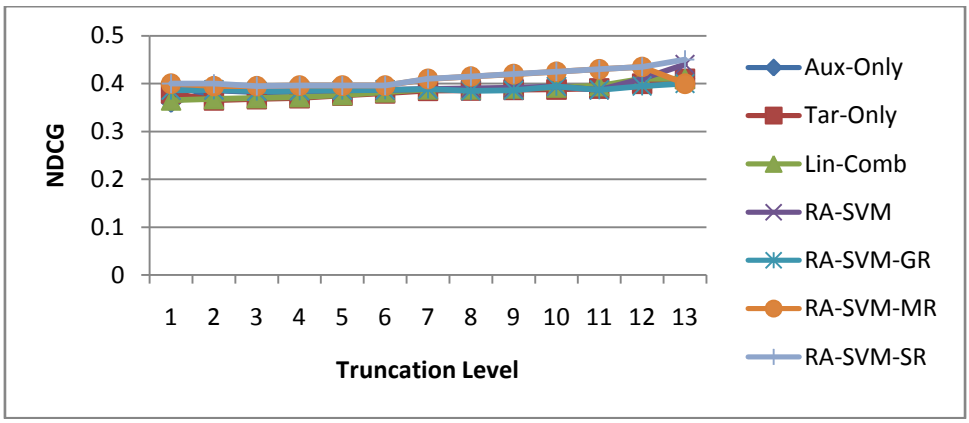

Fig. 5 – NDCG Results of web page search to image search adaptation with twenty labeled queries

As seen in fig. 5, adaptation performance is compared. The comparison is truncation level vs. NDCG. The proposed system outperforms the existing systems.

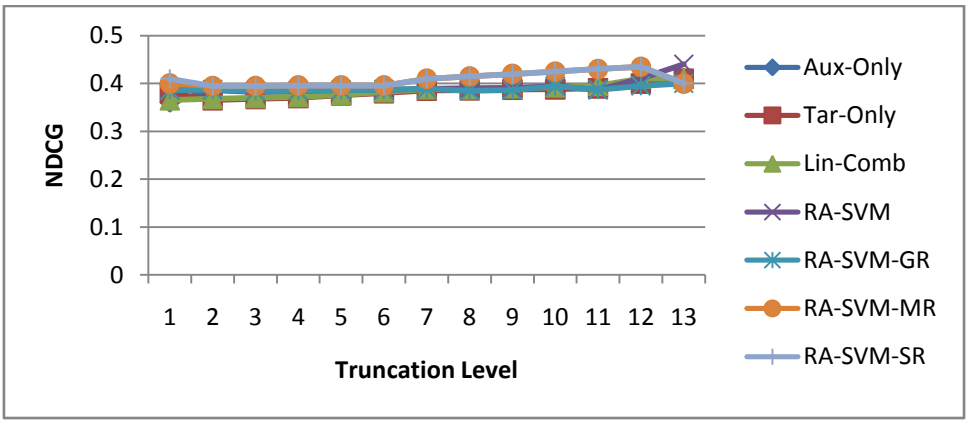

Fig. 6 – NDCG Results of web page search to image search adaptation with thirty labeled queries

As seen in fig. 6, adaptation performance is compared. The comparison is truncation level vs. NDCG. The proposed system outperforms the existing systems.

V. Conclusion

In this paper we have built a prototype application that is to demonstrate the efficiency of the ranking model that can adapt to new models. As of now the vertical search engines became popular. Developing different ranking models is not feasible. It will incur more cost and time. For this reason this paper proposed a new ranking model that can adapt to new domains easily. The proposed model works for various domains. For instance it can be adapted to medical search, educational search, vehicles search and so on. The prototype application has been tested and the experimental results revealed that the ranking model is adaptable to new domains.

References

[1] C.J.C. Burges, R. Reg.No, and Q.V. Le, "Learning to Rank withNo smooth Cost Functions," Proc. Advances in Neural InformationProcessing Systems (NIPS '06), pp. 193-200, 2006.

[2] Z. Cao and T. Yan Liu, "Learning to Rank: From PairwiseApproach to Listwise Approach," Proc. 24th Int'l Conf. MachineLearning (ICML '07), pp. 129-136, 2007.

[3] C.J.C. Burges, T. Shaked, E. Crenshaw, A. Lazier, M. Deeds, N.Hamilton, and G. Hullender, "Learning to Rank Using GradientDescent," Proc. 22th Int'l Conf. Machine Learning (ICML '05), 2005.

[4] Y. Freund, R. Iyer, R.E. Schapire, Y. Singer, and G. Dietterich, "An Efficient Boosting Algorithm for Combining Preferences,"J. Machine Learning Research, vol. 4, pp. 933-969, 2003.

[5] R. Herbrich, T. Graepel, and K. Obermayer, "Large Margin RankBoundaries for Ordinal Regression," Advances in Large MarginClassifiers, pp. 115-132, MIT Press, 2000.

[6] T. Joachims, "Optimizing Search Engines Using ClickthroughData," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discoveryand Data Mining (KDD '02), pp. 133-142, 2002.

[7] J. Blitzer, R. Mcdonald, and F. Pereira, "Domain Adaptation withStructural Correspondence Learning," Proc. Conf. Empirical Methodsin Natural Language Processing (EMNLP '06), pp. 120-128, July2006.

[8] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for TransferLearning," Proc. 24th Int'l Conf. Machine Learning (ICML '07),pp. 193-200, 2007.

[9] H. Shimodaira, "Improving Predictive Inference Under CovariateShift by Weighting the Log-Likelihood Function," J. StatisticalPlanning and Inference, vol. 90, no. 18, pp. 227-244, 2000.

[10] J. Yang, R. Yan, and A.G. Hauptmann, "Cross-Domain VideoConcept Detection Using Adaptive Svms," Proc. 15th Int'l Conf.Multimedia, pp. 188-197, 2007.

[11] B. Zadrozny, "Learning and Evaluating Classifiers Under SampleSelection Bias," Proc. 21st Int'l Conf. Machine Learning (ICML '04),p. 114, 2004.

[12] R. Klinkenberg and T. Joachims, "Detecting Concept Drift withSupport Vector Machines," Proc. 17th Int'l Conf. Machine Learning(ICML '00), pp. 487-494, 2000.

[13] J. Lafferty and C. Zhai, "Document Language Models, Query Models, and Risk Minimization for Information Retrieval," Proc. 24th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '01), pp. 111-119, 2001.

[14] J.M. Ponte and W.B. Croft, "A Language Modeling Approach to Information Retrieval," Proc. 21st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 275-281, 1998. GENG ET AL.: RANKING MODEL ADAPTATION FOR DOMAIN-SPECIFIC SEARCH 757

[15] S. Robertson and D.A. Hull, "The Trec-9 Filtering Track Final Report," Proc. Ninth Text Retrieval Conf., pp. 25-40, 2000.

[16] H. Daume III and D. Marcu, "Domain Adaptation for Statistical Classifiers," J. Artificial Intelligence Research, vol. 26, pp. 101-126, 2006.

[17] J.M. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "The Web as a Graph: Measurements, Models and Methods," Proc. Int'l Conf. Combinatorics and Computing, pp. 1-18, 1999.

[18] L. Page, S. Brin, R. Motwani, and T. Winograd, "The Pagerank Citation Ranking: Bringing Order to the Web," technical report, Stanford Univ., 1998.

[19] V.N. Vapnik, Statistical Learning Theory. Wiley-Interscience, 1998.

[20] F. Girosi, M. Jones, and T. Poggio, "Regularization Theory and Neural Networks Architectures," Neural Computation, vol. 7, pp. 219-269, 1995.

AUTHORS

Raja Rajeswari is student of Sri Sunflower College of Engineering and Technology, Hyderabad, AP, INDIA. She has received B.Tech Degree computer science and engineering, M.Tech Degree in computer science and engineering. Her main research interest includes data mining, Networking and Security.

Ganesh Datt is working as an Assistant Professor in Sri Sunflower College of Engineering and Technology, JNTUK, Hyderabad, Andhra Pradesh, India. He is pursuing Ph.D in Information Security. He has completed M.Tech (C.S.E) from JNTUH. His main research interest includes Cloud Computing and Data Mining

Sasi Krishna is student of Sri Sunflower College of Engineering and Technology, Hyderabad, AP, INDIA. She has received B.Tech Degree computer science and engineering, M.Tech Degree in computer science and engineering. Her main research interest includes data mining, Networking and Security.

