Design And Implementation of A cost Effective Ranking Adaptation Algorithm

K.Chiranjeevi, K.Archana

Dept of CSE M.Tech MLR Institute of Technology, Hyderabad, Andhra Pradesh, India
Dept of CSE Associate Professor in MLR Institute of Technology, Hyderabad, Andhra Pradesh, India

Abstract
Ranking plays an important role in vertical search domains as it helps users to view the best results quickly. This is required as search engines return huge number of records. Generally a ranking model is required for every domain as the data in each domain is different. However, it is tedious task to develop separate ranking model for each and every domain. A ranking model which can adapt to different domains can solve this problem. This paper proposes a new algorithm which adapts to various domains thus eliminating repetition of writing separate algorithm for each domain. An algorithm adapting to new domain reduces training cost thus making it cost-effective. We also proposed a ranking adaptability measurement for estimating the adaptability of ranking model. A prototype application is built to test the effectiveness of the application. The empirical results revealed that the proposed ranking model adaptation algorithm is capable of adapting to new domains.

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

1. Introduction

Information over Internet is being increased drastically every year. Users of Internet are able access required information on demand. The information retrieval systems generally return huge amount of data to end user. This causes the user's time to be wasted. In order to overcome this problem, ranking models have been around for many years as they are required by applications as part of information retrieval. While learning ranking model, documents are labeled based on their relevancies to the queries issued by users. There are many machine learning methods used for ranking. They include LambdaRank [1], ListNet [2], RankNet [3], RankBoost [4], and Ranking SVM [5], [6]. These kinds of algorithm have already shown their performance in information retrieval in the context of web search engines. There are domain specific search engines that have moved from broad based search to domain specific search known as vertical search for bestowing required information. The vertical search engines serve specific documents and document types. For instance a medical search engine only can retrieve information pertaining to medical field. In this manner there are different search engines for music, video, and images. They operate on different documents and their types or formats. The broad based search engines use ranking model that uses text search techniques. They even treat images as text based documents as they can use text associated with images. Some techniques use Term Frequency (TF) information for ranking documents. However, the broad based ranking model is generally built on multiple domains and which can't be generalized to work on specific search intentions of the user. Such model only can use ranking features of vertical domains. Semantic search is also kept in such ranking model. The search word is not considered as it is. Instead, it makes use of related and similar meaning words as well using lexical analysis. Building a ranking model for every domain is very tedious task. Therefore it is essential to have a model adaptation algorithm that can be used to switch from different domains and have search queries specific to the given domain. From experiments it is understood that the broad based search can give information that is reasonably useful though it may not be perfect. However, the broad based learning model can be adapted to various domain specific searches provided good labeling to documents in the domains. Thus it is well known that some prior knowledge about the domain information is required by the broad based ranking model to succeed in information retrieval. This appears to be costly and needs further improvement. Developing a ranking model that can adapt to new domains is very essential.

Ranking adaptation to new domains can be compared to classifier adaptation. There was some research in this area [7], [8], [9], [10], [11]. However, adaptation for the ranking problem is new and there is no solution found in the literature. However, classifier adaptation and concept drifting [12] existed. Classifier adaptation deals with binary targets while the ranking adaption is pertaining to predicting rankings on a corpus of documents. There are some problem with multiple classifiers and classifier adaptation which include the dependency of preference of documents, and difference in relevancy levels among the domains. This paper, instead of using labeled data, the focus is on developing a ranking model which can adapt to domain specific search. Model adaptation is important

Manuscript received August 5, 2015
Manuscript revised August 20, 2015

when compared to data adaptation or classifier adaptation. The problems of ranking model adaptation are going to overcome in this paper such as using labeled data of existing models, effective adaptation of ranking models and utilization of domain – specific features in the model adaptation. This paper addresses all these problems by developing a ranking adaptation algorithm that can adapt to various domain specific searches. It is a kind of black box adaptation model which needs some inputs so as to adapt to new domains.

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 experiments, results and evaluation while section 5 provides conclusions.

2. Related Work

There are many research works found in the literature. However, we present here works that are closely relevant to our paper which is meant for developing a new ranking model adaption algorithm for domain specific searches. There were many models presented in the literature for ranking documents that have been used by real world search engines. For instance Language Models for Information Retrieval [13], [14] and Classical BM25 [15] work well for broad based searches. These models work fine with few parameters to be adjusted. However, there are scenarios where different leaning methods, labeled data with complicated features that make the existing broad based ranking models to lag behind. For this reason we felt that there is a need for developing a new ranking model adaptation algorithm that can effectively adapt to new domains for best performance. Recently many ranking algorithms came into existence. They learn to rank the documents and they are based on machine learning techniques. Some of them convert the ranking problem into classification problem that act on pairs of documents with labeling. Examples for such algorithms include LambdaRank [1], ListNet [2], RankNet [3], RankBoost [4], and Ranking SVM [5], [6] etc. They all focus on the objective evaluation optimization for ranking. In this paper we are not going to develop a new learning algorithm as such. Instead we focus on adaptation of ranking models suitable for different domains that are based on the already available learn to rank kind of algorithms. Various adaptation methods came into existence that makes use of different classifiers. A mixture model was presented by Daume and Marcu in order to address problems pertaining to domain distribution differences between testing sets and training sets [16]. For the same a boosting model was presented in [8]. A structural correspondence learning method was introduced by Blitzer et al. [7]. It is used to mine different domains in terms of correspondences of features. Yang et al. [10] proposed an algorithm for detecting cross domain videos. The algorithm is named as Adaptive SVM. All these works specified here are meant for designing classification problems. However, this paper focuses on ranking model adaption for various domains.

3. Raking Adaptation

This section provides information about the proposed ranking adaption model. First of all we describe the adaptation problem here. We consider a set of queries represented by Q, set of documents represented by D. The search results are labeled by human annotators. Other ranking information considered here include HITS [17] and PageRank [18] etc. The purpose of learning to rank algorithms is to estimate ranking. We assume that the number of returned documents and the number of queries in the training set to be small. As auxiliary ranking models have prior knowledge on labeled dataset, they need few training samples in order to adapt the ranking model.

Ranking Adaptation SVM

We assume that when the target domain and the auxiliary domain are similar, it is possible that ranking functions for them also tend to be similar. With this assumption the prior knowledge on dataset can be used in the ranking adaption model. The conventional regularization frameworks such as SVM [19] and neural networks [20] have some problem. The problem is known as ill-posed problem that which contains prior assumption and also data. This problem can be elegantly solved using a regularization framework that makes use of respective adaptive ranking function. The proposed ranking adaption is formulated as:

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

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

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

Adapting to Multiple Domains

The proposed algorithm can be extended to support multiple domains seamlessly in which ranking models can be learned from multiple domains in order to facilitate domain specific search with a common ranking adaption model. Assuming that there are certain auxiliary functions which can be used in the proposed ranking model, the multiple domain adaptions can be formulated as:

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

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

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

Generally data that comes from different domains should have various features specific to domain from which data comes. The proposed ranking model adapts to such features automatically. The utilization of domain specific features into the ranking adaptation model is essential and the proposed model is using such information. The ranking loss concept is also considered in the proposed system and the consistency constraint has been incorporated. Based on the similarity between the documents which are being processed, rescaling degree is controlled by looking that the document similarities. For making loss for pair wise documents, slack scaling is used. The margin rescaling is an optimization problem that rescales margin violations of adaptability. The margin rescaling is formulated as follows:

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

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

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

In the same fashion slack rescaling is formulated as follows:

$$\text{Max} - 1/2 \sum_{ijkk}^{n} \quad \sum_{ijk}^{n} \quad \varpi ijk\ \varpi lmn\ X^T_{ijk} X_{lmn}$$

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

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

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

4. Experimental Results

The experiments are carried out using a prototype web application. The application was built in Java/J2EE platform which make use of JDBC (Java Database Connectivity), Servlets and JSP (Java Server Pages). A PC with 2 GB RAM with Pentium Core 2 Dual processor is used for experiments. With regard to datasets, TD2003 and TD2004 benchmark datasets are gathered. The performance of the proposed ranking adaptation model is evaluated by using two measures namely normalized discounted cumulative gain and mean average precision. The results of the proposed method are compared many other baseline methods including Aux-Only, Tar-Only, and Lin-Comb.

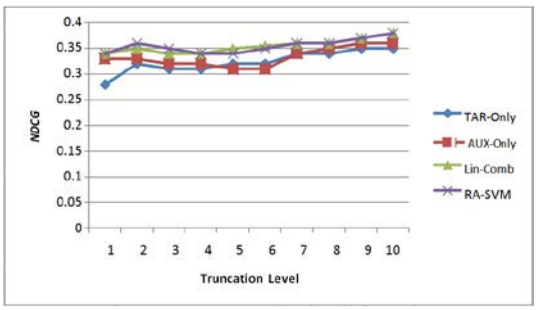

Fig. 1 - TD2003 to TD2004 adaptation with five queries

As can be seen in fig. 1, comparison is made with adaptation performance of proposed algorithm with other three algorithms. The adaptation is from TD2003 dataset to TD2004 dataset with five queries. Out of all, the proposed algorithm has shown best performance. When Aux-Only is compared with Tar-Only, the Aux-Only outperforms the other model.

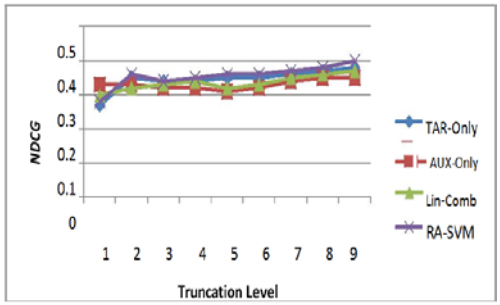

Fig. 2 - TD2003 to TD2004 adaptation with ten queries

As can be seen in fig. 2, comparison is made with adaptation performance of proposed algorithm with other three algorithms. The adaptation is from TD2003 dataset to TD2004 dataset with ten queries. Out of all, the proposed algorithm has shown best performance. As number of queries is increased, when Aux-Only is compared with Tar-Only, the Tar-Only outperforms the other model.

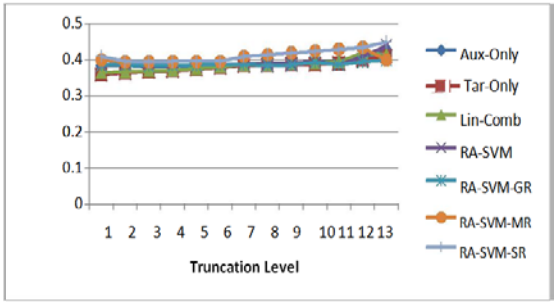

Fig. 3 – NDCG Results of web page search to image search adaptation with five labeled queries

As can be seen in fig. 3, adaptation is made from web page search to image search with five labeled queries. The performance of proposed model is compared with other baseline models. The proposed model outperforms other models.

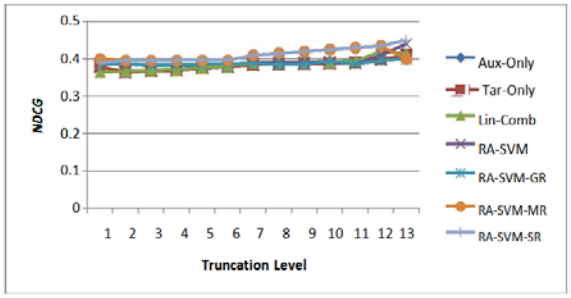

Fig. 4 – NDCG Results of web page search to image search adaptation with ten labeled queries

As can be seen in fig. 4, adaptation is made from web page search to image search with ten labeled queries. The performance of proposed model is compared with other baseline models. The proposed model outperforms other models.

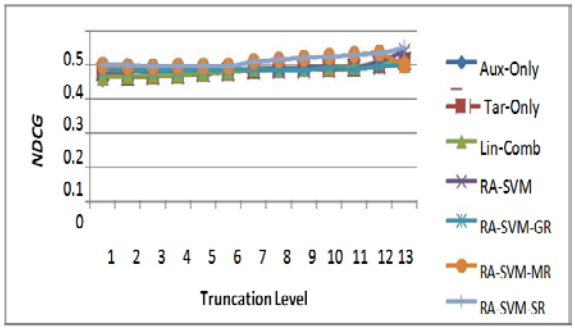

Fig. 5 – NDCG Results of web page search to image search adaptation with twenty labeled queries

As can be seen in fig. 5, adaptation is made from web page search to image search with twenty labeled queries. The performance of proposed model is compared with other baseline models. The proposed model outperforms other models.

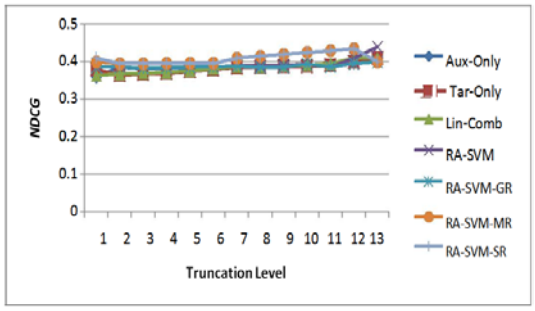

Fig. 6 – NDCG Results of web page search to image search adaptation with thirty labeled queries

As can be seen in fig. 6, adaptation is made from web page search to image search with thirty labeled queries. The performance of proposed model is compared with other baseline models. The proposed model outperforms other models.

Conclusion

This paper proposed a new ranking model adaption algorithm based on existing models that can adapt to various domains. As vertical search engines became popular, they can provide only information specific to a domain. The problem with such ranking models used by vertical search engines is that each model works only for one domain. This is because; data and data types of each domain are different. For instance medical search engine handles only medical data while vehicles search engine can only work for that domain. This is the motivation behind the work in this paper. We proposed a ranking model adaption algorithm for ranking search results of various domains. We built a prototype web application for testing the efficiency of the proposed ranking model. The experimental results revealed that the proposed model is adaptable and useful for domain specific search.

References

[1] C.J.C. Burges, R. Ragno, and Q.V. Le, “Learning to Rank with Nonsmooth Cost Functions,” Proc. Advances in Neural Information Processing Systems (NIPS ’06), pp. 193-200, 2006.

[2] Z. Cao and T. Yan Liu, “Learning to Rank: From Pairwise Approach to Listwise Approach,” Proc. 24th Int’l Conf. Machine Learning (ICML ’07), pp. 129-136, 2007.

[3] C.J.C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, “Learning to Rank Using Gradient Descent,” 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 Rank Boundaries for Ordinal Regression,” Advances in Large Margin Classifiers, pp. 115-132, MIT Press, 2000.

[6] T. Joachims, “Optimizing Search Engines Using Clickthrough Data,” Proc. Eighth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD ’02), pp. 133-142, 2002.

[7] J. Blitzer, R. Mcdonald, and F. Pereira, “Domain Adaptation with Structural Correspondence Learning,” Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP ’06), pp. 120-128, July 2006.

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

[9] H. Shimodaira, “Improving Predictive Inference Under Covariate Shift by Weighting the Log-Likelihood Function,”

J. Statistical Planning and Inference, vol. 90, no. 18, pp. 227-244, 2000.
[10] J. Yang, R. Yan, and A.G. Hauptmann, "Cross-Domain Video Concept Detection Using Adaptive Svms," Proc. 15th Int'l Conf. Multimedia, pp. 188-197, 2007.
[11] B. Zadrozny, "Learning and Evaluating Classifiers Under Sample Selection Bias," Proc. 21st Int'l Conf. Machine Learning (ICML '04), p. 114, 2004.
[12] R. Klinkenberg and T. Joachims, "Detecting Concept Drift with Support 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.

K.Chiranjeevi has received B.Tech from Pragati Engineering college. and pursuing M.Tech (C.S.E) in M.L.R.Institute of Technology,, JNTUH, Hyderabad, Andhra Pradesh, India.His main research interest includes Data Mining, Information Security, network protection and security control.

K.ARCHANA completed my M.Tech in Computer Science at Jawaharlal Nehru Technological University Hyderabad. I have 7 Yrs, experience in teaching. Presently am working in MLR Institute of Technology, Hyderabad as a Associate Professor in Department of Computer Science and Engineering. My specialized area is Network security and Image processing. I published 3 papers in that area ,two are International and one national(Conference).