Ranking Model for Domain Specific Search

Priyanka Jadhav[1] Vaishali Pawar[2] Chaitali Jadhav[3] Prof. Nidhi Sharma[4]
[1,2,3]Student [4]Assistant Professor
[1,2,3,4]Department of Computer Engineering
[1,2,3,4]Bharati Vidyapeeth College of Engineering, Navi Mumbai-400614

Abstract— Learning to rank is an important area at the interface of machine learning, information retrieval and Web search. The technology has been successfully applied to web search, and is becoming one of the key machines for building search engines. The central challenge in optimizing various measures of ranking loss is that the objectives tend to be non-convex and discontinuous. In recent years, boosting, neural networks, support vector machines, and other techniques have been applied. To build a unique ranking model for each domain it time-consuming for training models. In this paper, we address these difficulties by proposing algorithm called ranking adaptation SVM (RA-SVM).Our algorithm only requires the prediction from the existing ranking models, rather than their internal representations or the data from auxiliary domains.

Key words: Learning to rank, RA-SVM, Ranking of Candidates

I. INTRODUCTION

Search engines are programs that search documents for specified keywords and return a list of the documents where the keywords were found. By typing in a word or phrase (known as a keyword), the search engine will produce pages of links on that topic. The more relevant links are at the top of the list, but that is not always true. The information contains in the search engines is may be specialist in web pages, images and other types of files. The learning to rank is a kind of learning based information restoration techniques, specialized in the learning a ranking model with some documents labeled with their relevancies to some queries.

Page Rank is a link analysis algorithm applied by Google.com that assigns a number or rank to each hyperlinked web page within the World Wide Web. The basic purpose of PageRank is to list web pages from the most important to the least important, reflecting on a search engine results page when a keyword search occurs. The basic process involves PageRank evaluating all of the links to a particular web page [4]. If a web page has a lot of links from large websites that also rank well, then the original web page is given a high ranking. Where the links are coming from is just as important as the number of links to any particular web page, the system being rather "democratic" according to Google.com.

Domain Specific Search focus on one area of knowledge, creating customized search experiences, that because of the domain's limited corpus and clear relationships between concepts, provide extremely relevant results for searchers. However, as the emergence of domain-specific search engines [4], more attentions have moved from the broad based search to specific verticals, for hunting information constraint to a certain domain. Different vertical search engines deal with different topicalities, document types or domain-specific features. For example, a job search engine should clearly be specialized in terms of its topical focus, whereas a music, image or video search engine would concern only the documents in particular formats.

In this paper, the adaptation of ranking models is focused upon, as a substitute of using the labeled data from auxiliary domains directly, which could not be accessible due to missing data or privacy issue. Moreover, the Model adaptation is more advantageous and efficient than data adaptation, because the learning complexity is correlated with the size of the training set of the target domain, which is much smaller than the size of auxiliary training data set that is used for ranking adaptation. Such a ranking model adaptation is much efficient.

II. RANKING SVM

There are many services provided by the ranking models are effectively some issues are risen in these model. The general difficulties faced by the classifier adaption namely: covariate shift and concept drifting and it have more challenging compared to the ranking models. The classifier adaptation, which mainly deals with binary targets, ranking adaptation desires to adapt the model which is used to predict the rankings for a collection of documents. Though the documents are normally labeled with several relevance levels, which seem to be able to be handled by a multiclass classification or regression, it is still difficult to directly use classifier adaption for ranking. The reason lies in two ways: 1) in ranking, the mainly concentrated is about the preference of two documents or the ranking of a collection of documents, which is difficult to be modeled by classification or regression; 2) the relevance levels between different domains are sometimes different and need to be aligned.

SVM is supervised learning methods that analyze data and recognize patterns, used for classification. Similar to the basic SVM, the motivation of Ranking SVM [2] is to identify a one dimensional linear subspace, based on some criteria the points can be ordered into the optimal ranking list under some criteria. Thus, the ranking function used in the adaptation takes the structure of the linear model f(Ó(q,d))=WTÓ(q,d), here the bias parameter is uncared for, because the ranking list that is produced finally is sorted by the prediction f and is invariant to the bias. Ranking SVM is the base of this paper.

III. OPTIMIZATION PROBLEM

The optimization problem for Ranking SVM is defined as follows:

$$\min_{f,\xi_{ijk}} \frac{1}{2}||f||^2 + C\sum_{i,j,k}\xi_{ijk}$$
$$\text{s.t.} \quad f(\phi(q_i,d_{ij})) - f(\phi(q_i,d_{ik})) \geq 1-\xi_{ijk}$$
$$\xi_{ijk} \geq 0,$$
$$\text{for} \quad \forall i \in \{1,2,\ldots,M\},$$
$$\forall j \forall k \in \{1,2,\ldots,n(q_i)\} \text{ with } y_{ij} > y_{ik},$$

Where C is the trade-off parameter for balancing the large-margin regularization ||f ||2 and the loss term □i,j,k ξijk Because f is a linear model , we can derive that

f (φ(qi , dij))−f (φ(qi , dik)) = f (φ(qi , dij)−φ(qi , dik)), with φ(qi , dij)−φ(qi , dik) denoting the difference of the feature vectors between the document pair dij and dik . If we further introduce the binary label sign(yij − yik) for each pair of documents dij and dik , the above Ranking SVM problem can be viewed as a standard SVM for classifying document pairs into positive or negative, i.e., whether the document dij should be ranked above dik or not.

IV. RANKING ADAPTATION WITH DOMAIN-SPECIFIC FEATURE

Conventionally, data from different domains are also characterized by some domain-specific features, e.g., when we adopt the ranking model learned from the webpage search domain to the images search domain the image content can provide additional information to facilitate the text-based ranking model adaptation. In this section, we discuss how to utilize these domain-specific features, which are usually difficult to translate to textual representations directly, to further boost the performance of the proposed RA-SVM. The basic idea of our method is to assume that documents with similar domain-specific features should be assigned with similar ranking predictions. We name the above assumption as the consistency assumption, which implies that a robust textual ranking function should perform relevance prediction that is consistent to the domain-specific features. To implement the consistency assumption, we are inspired by the work [4], the ranking loss is directly correlated with the slack variable, which stands for the ranking loss for pairwise documents, and is nonzero as long as the ranking function predicts a wrong order for the two documents. In addition, as a large margin machine, the ranking loss of RA-SVM is also correlated with the large margin specified to the learned ranker. Therefore, to incorporate the consistency constraint, we rescale the ranking loss based on two strategies, namely margin rescaling and slack rescaling. The rescaling degree is controlled by the similarity between the documents in the domain-specific feature space, so that similar documents bring about less ranking loss if they are ranked in a wrong order.

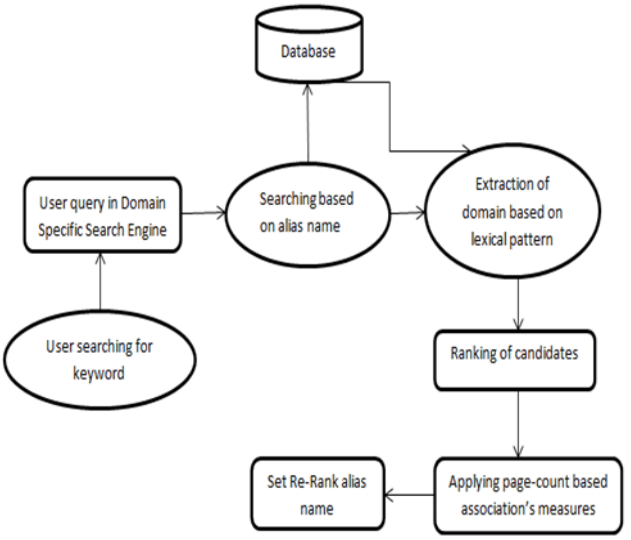

Fig. 1: Ranking Adaptation with Domain-Specific Feature

- Ranking of Candidates
- Applying Page-Count-Based Association Measures
- Applying Robust Alias Extraction and detection system

A. Ranking Of Candidates:

Considering the noise in web snippets, candidates extracted by the shallow lexical patterns might include some invalid aliases. From among these candidates, we must identify those, which are most likely to be correct aliases of a given name. We model this problem of alias recognition as one of ranking candidates with respect to a given name such that the candidates, who are most likely to be correct aliases are assigned a higher rank. First, we define various ranking scores to measure the association between a name and a candidate alias using three different approaches: lexical pattern frequency, word co-occurrences in an anchor text graph and page counts on the web.

B. Applying Page-Count-Based Association Measures:

We defined various ranking scores using anchor texts. However, not all names and aliases are equally well represented in anchor texts. Consequently, in this section, we define word association measures that consider co-occurrences not only in anchor texts but in the web overall. Page counts retrieved from a web search engine for the conjunctive query, "p and x," for a name p and a candidate alias x can be regarded as an approximation of their co-occurrences in the web. We compute popular word association measures using page counts returned by a search engine.

C. Applying Robust Alias Extraction And Detection System:

We compare the proposed SVM-based method against various individual ranking scores (baselines) and previous studies of alias extraction on Japanese personal names data set Overall, the proposed method extracts most aliases in the manually created gold standard (shown in bold). It is noteworthy that most aliases do not share any words with the name nor acronyms, thus would not be correctly extracted from approximate string matching methods. Following earlier research on web-based social network extraction and we measured the association between two people using the PMI between their names on the web.

V. PROCEDURE

This paper is integrated with following Modules:

- Ranking Adaptation Module.
- Ranking adaptability Measurement Module.
- Domain specific ranking Model Module.
- Ranking Support Vector Machine Module.

A. Ranking Adaptation Module:

A ranking-model adaptation module is used to adapt a ranking model utilized by a general domain for use with a specific domain resulting in an adapted ranking model. Ranking adaptation is closely related to classifier adaptation. Unlike classifier adaptation, which mainly deals with binary targets, ranking adaptation desires to adapt the model which is used to predict the rankings for a collection of domains. In

ranking the relevance levels between different domains are sometimes different and need to be aligned. Ranking models is adapted for the existing broad-based search or some verticals, to a new domain, so that the amount of labeled data in the target domain is reduced while the performance requirement is still guaranteed. The ranking-model adaptation module includes, without limitation, a ranking adaptation support vector machines (SVM) module and a ranking adaptability measurement module.

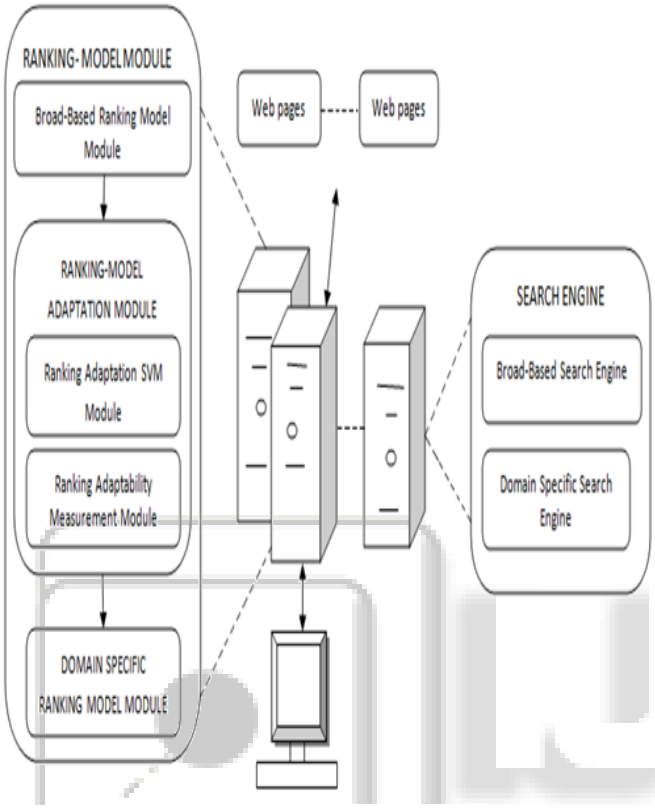

Fig. 2: Ranking model module

B. Ranking Adaptability Measurement Module:

A ranking adaptability measurement module to quantitatively estimate an adaptability of the first ranking model; and a quadratic program solver utilized to solve a quadratic optimization problem determined by the ranking-model adaptation module. The ranking adaptability measurement module investigates the correlation between two ranking lists of the labeled documents in the target domain.

C. Domain Specific Ranking Model Module:

If the user query is performed using domain-specific search engine, a ranking-model adaptation module enables the broad-based ranking-model module to be modified into a domain-specific ranking-model module. The domain-specific ranking-model module reduces the search results to the most relevant pages, with respect to the search terms input by the user into the domain-specific search engine. The basic idea of our method is to assume that documents with similar domain-specific features should be assigned with similar ranking predictions. We name the above assumption as the consistency assumption, which implies that a robust textual ranking function should perform relevance prediction that is consistent to the domain-specific features.

D. Ranking Support Vector Machine Module:

The ranking-model adaptation module utilizes a ranking adaptation support vector machines (SVM) module to perform the adaptation. Another advantage of utilizing the adapted ranking SVM is referred to as the black-box adaptation. Ranking Support Vector Machines (Ranking SVM), which is one of the most effective learning to rank algorithms, and is here employed as the basis of our proposed algorithm, the proposed RA-SVM does not need the labeled training samples from the auxiliary domain, but only its ranking model f^a.

VI. Experimental Work

In order to experimentally verify the ranking model Ranking Adaptation Support Vector Machine, the following has been done.

A. Domain Creation:

A domain has been created that contains the information for the user needs. This domain is maintained by the administrator. This is more like a domain specific search engine. Job domain has details regarding certain specific job information.

B. Search Page:

The user will be allowed to enter the queries in the search page. Based on the entered query, information will be retrieved to the user.

C. Page Ranking:

The retrieved page will be brought to the user. The most viewed page will be given the first rank. Ranking will be done based on the user views. If a page is viewed many times, that will be given the first priority. As the user views a page, the page count increases making it higher than the other pages. For experimental verification, a domain search engine with job domain is created. The user is allowed to search for a keyword "job". Then under when nokari.com is searched for the most viewed pages is ranked first using RA-SVM.

The following graph shows that the nokari.com is viewed the maximum times and is ranked higher than any other pages.

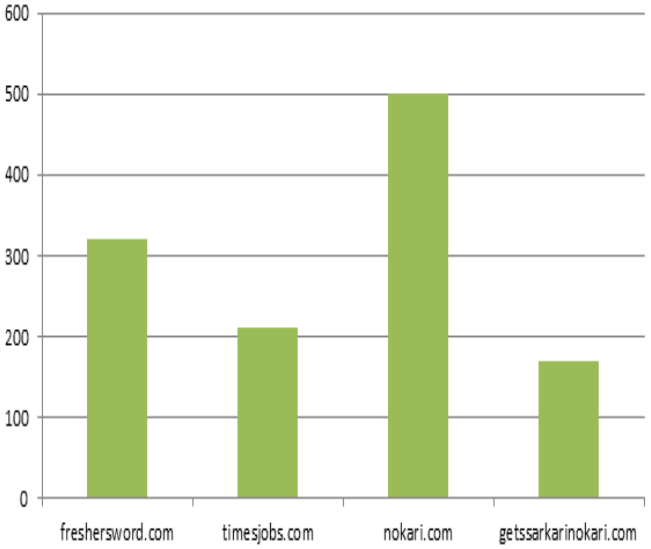

Fig. 3: Graph

The RA-SVM [4] has several merits, which makes the algorithm more flexible to be applied practically.

Model adaptation: The Ranking Adaptation SVM do not consider the labeling of training samples but it takes into account only the ranking model f^a. This kind of model adaptation is meritorious than data-based adaptation since the data might be missing or unavailable due to privacy issue. The ranking model is very easy to access.

Black-box adaptation: Another advantage of utilizing the adapted ranking SVM is referred to as the black-box adaptation. For example, to obtain the ranking adaptation SVM, the internal representation of the model f^a not needed. Only the prediction of the auxiliary model is used in conjunction with the training samples of the target domain. Utilizing the black-box adaptation eliminates the necessity to know what learning model (algorithm) the auxiliary domain is using to predict results. Again, only the ranking function model predictions are necessary

Reducing the labeling cost: Only a very few number of samples are to be labeled while adapting the auxiliary ranking model. Hence, the labeling cost is reduced. Reducing the computational cost: The ranking adaptation SVM algorithm could be transformed into a Quadratic Programming problem. In this the learning complexity can be directly related to the number of labeled samples that are taken under consideration in the new or target domain.

VII. FURTURE ENHANCEMENT

Every application has its own merits and demerits. The project has covered almost all the requirements. Further requirements and improvements can easily be done since the coding is mainly structured or modular in nature. Changing the existing modules or adding new modules can append improvements. Further enhancements can be implemented in this project. Since this project is concerned with a specific domain "languages" it can be further extended to various domains. Image search, document retrieval, map search can also be implemented in this.

VIII. CONCLUSION

As increases in various vertical search engines, building one model for each vertical domain is both laborious for labeling the data and time-consuming for learning the model. In this paper, we propose the ranking model adaptation, to adapt the well learned models from the broad-based search or any other auxiliary domains to a new target domain. With this approach many users will get benefited and they will get the expected results in less time as the algorithm implemented will apply the ranking function and will give the highest rank allotted data for the user specific search.

By model adaptation, only a small number of samples need to be labeled, and the computational cost for the training process is greatly reduced. Based on the regularization framework, the Ranking Adaptation SVM algorithm is proposed, which performs adaptation in a black-box way and only relevance predication of the auxiliary ranking models is needed for the adaptation.

IX. ACKNOWLEDGEMENT

Sincere appreciation and warmest thanks are extended to the many individuals who is in their own ways have inspired us in the completion of this project.

Firstly we are thankful to our principal DR. M. Z. Shaikh for his help. We are extremely grateful for his friendly support and professionalism. We express our heartfelt gratitude to our Head of Department Prof. D. R. Ingle & project coordinator Prof. Rahul Patil of Computer Department for their help and support. This task would have not been possible without the help and guidance of our esteemed project supervisor Prof.Nidhi Sharma, without her expert help and guidance, this project would not have reached this stage. We are also conveying special thanks to all staff members of Computer Engineering Department for their support and help. Last but not least, we are very much thankful to our friends who directly or indirectly helped us in completion of the project report.

X. REFERENCES

[1] Blitzer. J, Mcdonald. D, Pereira. R, (July 2006) "Domain Adaptation with Structural Correspondence Learning", Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP '06), pp. 120-128.
[2] Bo Geng, Linjun Yang, Chao Xu, Xian-Sheng Hua, (Apr 26,2012) "Ranking Model Adaptation for Domain-Specific Search", United States Patent Publications.
[3] Chapelle. O, Keerthi. S, (July 20, 2009)"Efficient Algorithms for Ranking with SVMs", Information Retrieval Journal.
[4] Geng. B, Yang. L, Xu. C, and Hua. X, (2009) "Ranking Model Adaptation for Domain-Specific Search", Proc. 18th ACM Conf. Information and Knowledge Management (CIKM '09), pp. 197206.
[5] H. Daume III and D. Marcu, "Domain Adaptation for Statistical Classifiers," J. Artificial Intelligence Research, vol. 26, pp. 101-126, 2006.
[6] 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. 933969, 2003.
[7] 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.
[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] 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.
[10] 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.
[11] Matteo Pasquinelli,"Google's PageRank Algorithm: A Diagram of the Cognitive Capitalism and the Rentier of the Common Intellect".

[12] Ming-Feng Tsai, Tie-Yan Liu, (2000) ”FRank: A Ranking Method with Fidelity Loss”, SIGIR’07, July 23–27, Amsterdam, The Netherlands.
[13] Ponte. J. M and Croft. W. B, (1998)“A Language Modeling Approach to Information Retrieval”, Proc. 21st Ann. Int’l ACM SIGIR Conf. Research.

