Adapting a Ranking Model for Domain-Specific Search

M.Sirisha [*], Sanjeeva Rao Sanku [**]

[*] Pursuing M.Tech in Computer Science and Engineering from Nagole Institute of Technology and Science, JNTUH, A.P, INDIA.
[**] Assoc. Prof. Dept. CSE, M.Tech,CSE(from JNTU Kakinada

Abstract- An adaptation process is described to adapt a ranking model constructed for a broad-based search engine for use with a domain-specific ranking model. It's difficult to applying the broad-based ranking model directly to different domains due to domain differences, 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 The ranking model is adapted for use in a search environment focusing on a specific segment of online content, for example, a specific topic, media type, or genre of content. a domain-specific ranking model reduces search results to the data from a specific domain that are relevant with respect to the search terms input by the user. The ranking order may be determined with reference to a given numerical score, an ordinal score, or a binary judgment such as "relevant" or "irrelevant".

Index Terms- broad based search, Domain Adaptation, Support Vector Machines.

I. INTRODUCTION

LEARNING to rank is a kind of learning based information retrieval techniques, specialized in learning a ranking model with some documents labelled with their relevancies to some queries, where the model is hopefully capable of ranking the documents returned to an arbitrary new query automatically. Based on various machine learning method, Ranking the learning to rank algorithms have already shown their promising performances in information retrieval, especially Web search. However, as the emergence of domain-specific search engines, 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 medical 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.

Since currently the broad-based and vertical search engines are mostly based on text search techniques, the ranking model learned for broad-based can be utilized directly to rank the documents for the verticals. For example, most of current image search engines only utilize the text information accompanying images as the ranking features, such as the term frequency (TF) of query word in image title, anchor text, alternative text, surrounding text, URL and so on. Therefore, Web images are actually treated as text-based documents that share similar ranking features as the document or Web page ranking, and text based ranking model can be applied here directly. However, the broad-based ranking model is built upon the data from multiple domains, and therefore cannot generalize well for a particular Domain with special search intentions.

II. RANKINGADAPTATION

We define the ranking adaptation problem formally as follows: for the target domain, a query set $Q = \{q1 , q2 , . . . , qM \}$ and a document set $D = \{d1 , d2 , . . . , dN \}$ are given. For each query $qi \in Q$, a list of documents $di = \{di1 , di2 , . . . , di,n(qi) \}$ are returned and labeled with the relevance degrees $yi = \{yi1 , yi2 , . . . , yi,n(qi) \}$ by hu- man annotators. The relevance degree is usually a real value, i.e., $yij \in R$, so that different returned documents can be compared for sorting an ordered list. For each query document pair $< qi , dij >$, an s-dimensional query dependent feature vector $\varphi(qi , dij) \in Rs$ is extracted, e.g., the term frequency of the query keyword qi in the title, body, URL of the document dij . Some other hyperlink based static rank information is also considered, such as Page rank, HITS and so on. n(qi) denotes the number of returned documents for query qi. The target of learning to rank is to estimate a ranking function $f \in Rs \rightarrow R$ so that the documents d can be ranked for a given query q according to the value of the prediction $f (\varphi(q, d))$. In the setting of the proposed ranking adaptation, both the number of queries m and the number of the returned documents n(qi) in the training set are assumed to be small. They are insufficient to learn an effective ranking model for the target domain. However, an auxiliary ranking model f a , which is well trained in another domain over the labeled data Qa and Da , is available. It is assumed that the auxiliary ranking model f a contains a lot of prior knowledge to rank documents, so it can be used to act as the base model to be adapted to the new domain. Few training samples can be sufficient to adapt the ranking model since the prior knowledge is available.Before the introduction of our proposed ranking adaptation algorithm, it's important to review the formulation of 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.

A. Ranking SVM

Similar to the conventional Support Vector Machines (SVM) for the classification problem,the motiva- tion of Ranking SVM is to discover a one dimensional linear subspace, where the points can be ordered into the optimal ranking list under some criteria. Thus, the ranking function takesthe form of the linear model $f (\varphi(q, d)) = wT \varphi(q, d)$, where the bias parameter is

ignored, because the final ranking list sorted by the prediction f is invariant to the bias. The optimization proble for Ranking SVM is defined as follows:

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

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. Since the number of labeled samples for the new domain is small, if we train the model using only the samples in the new domain, it will suffer from the insufficient training sample problem, which is ill-posed and the solution may be easily overfitting to the labelled samples with low generalization ability. Moreover, the current SVM solver requires super-quadratic computational cost for the training, as a consequence, it is quite time-consuming and nearly infeasible to train models using the training data from both the auxiliary domain and the target domain. This problem is more severe for the ranking SVM since the training are based on pairs and so the problem size is quadratic to the sample size. In the following, we will develop and algorithm to labelled in the new domain. By model adaption, both the effectiveness of the result ranking model and the efficiency of the training process are achieved.

B. Ranking Adaptation SVM

It can be assumed that, if the auxiliary domain and the target domain are related, their respective ranking func- tions f a and f should have similar shapes in the function space Rs → R. Under such an assumption, f a actually provides a prior knowledge for the distribution of f in its parameter space. The conventional regularization framework, such as Lp -norm regularization, manifold regularization designed for SVM,regularized neural network and so on, shows that the solution of an ill-posed problem can be approximated from variational principle, which contains both the data and the prior assumption.Consequently, we can adapt the regularization framework which utilizes the f a as the prior information, so that the ill-posed problem in the target domain, where only few query document pairs are labeled, can be solved elegantly. By modeling our assumption into the regularization term, the learning problem of Ranking Adaptation SVM (RA-SVM) can be formulated as:

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

The objective function (2) consists of the adaptation regularization term ‖f − f a‖2 , which minimizes the dis-tance between the target ranking function and the auxil-iary one in the function space or the parameter space, to make them close; the large-margin regularization ‖f ‖2 ; and the loss term Pi,j,k ξijk . The parameter δ ∈ [0, 1] is a trade-off term to balance the contributions of large- margin regularization ‖f ‖2 which makes the learned model numerically stable, and adaptation regularization ‖f − f a‖2 which makes the learned model similar to the auxiliary one. When δ = 0, Problem (2) degrades to the conventional Ranking SVM (1), in other words, RA-SVM is equivalent to directly learning Ranking SVM over the target domain, without the adaptation of f a. The parameter C is the same as in Ranking SVM, for balanc-ing the contributions between the loss function and the regularization terms. It can be observed that when C = 0 and δ = 1, Eq. (2) actually discards the labeled samples in the target domain, and directly output a ranking function with f = f a . This is sometimes desirable, since if the labeled samples in the target domain are unavailable or unusable, f a is believed to be better than random guess for ranking the documents in the target domain,as long as the auxiliary domain and the target domain are related.

C.Optimization Methods

To optimize Problem (2), we briefly denote xijk = φ(qi , dij) − φ(qi , dik) and introduce the Lagrange multi- pliers to integrate the constraints of (2) into the objective function, which results in the primal problem:

$$\begin{aligned} L_P = & \frac{1-\delta}{2}||f||^2 + \frac{\delta}{2}||f-f^a||^2 + C\sum_{i,j,k}\xi_{ijk} \\ & \vdash \sum_{i,j,k}\mu_{ijk}\xi_{ijk} - \sum_{i,j,k}\alpha_{ijk}(f(\mathbf{x}_{ijk}) - 1 + \xi_{ijk})). \end{aligned} \quad (3)$$

Taking the derivatives of LP w.r.t. f , and setting it to zero, we can obtain the solution as:

$$f(\mathbf{x}) = \delta f^a(\mathbf{x}) + \sum\nolimits_{i,j,k}\alpha_{ijk}\mathbf{x}_{ijk}^T\mathbf{x}. \quad (4)$$

Denoting Δf (x) = Σ i,j,k αijk xT Viewed as the part of support vectors learned from the target domain, we can derive from (4) that the final ranking function f , which we would like to achieve for the target domain, is a linear combination between the auxiliary function f a and the target part Δf , and the parameter δ controls the contribution of f a .In addition to (4), the optimal solution of problem (2) should satisfy the Karush-Kuhn-Tucker (KKT) con-ditions, which are composed of:

$\alpha ijk\ (f\ (xijk\ ) - 1 + \xi ijk\ ) = 0$
$\alpha ijk\ \geq 0$
$f\ (xijk\ ) - 1 + \xi ijk\ \geq 0$
$\mu ijk\ \xi ijk\ = 0$
$\mu ijk\ \geq 0$
$\xi ijk\ \geq 0$
$C - \alpha ijk\ - \mu ijk\ = 0.$ (5)

Substituting (4) and (5) back into (3), we can derive the dual problem formulation as:

$$\max_{\alpha_{ijk}} \ -\frac{1}{2}\sum_{i,j,k}\sum_{l,m,n}\alpha_{ijk}\alpha_{lmn}\mathbf{x}_{ijk}^{T}\mathbf{x}_{lmn} + \sum_{i,j,k}(1-\delta f^{a}(\mathbf{x}_{ijk}))\alpha_{ijk}$$
$$\text{s.t.}\ \ 0 \leq \alpha_{ijk} \leq C,$$
$$\text{for}\ \ \forall i \in \{1,2,\ldots,M\},$$
$$\forall j \forall k \in \{1,2,\ldots,n(q_i)\}\ \text{with}\ y_{ij} > y_{ik}. \qquad (6)$$

The above problem is a standard Quadratic Programming(QP) problem, and any standard QP solvers, e.g. over fitting problem can be overcome by utilizing the prior information from the auxiliary model.

D. Discussions

The proposed RA-SVM has several advantages, which makes our algorithm highly applicable and flexible when applied to the practical applications. We'll give more discussions of the characteristics of RA-SVM in the following.

- Model adaptation: the proposed RA-SVM does not need the labeled training samples from the auxil-iary domain, but only its ranking model f a . Such a method is more advantageous than data based adaptation, because the training data from auxiliary domain may be missing or unavailable, for the copy- right protection or privacy issue, but the ranking model is comparatively easier to obtain and access.
- Black-box adaptation: The internal representation of the model f a is not needed, but only the prediction of the auxiliary model to the training samples from the target domain f a(x) is used. It brings a lot of flexibilities in some situations where even the auxiliary model itself may be unavailable. Also, in some cases, we would like to use a more advanced algorithm for learning the ranking model for the new target domain, than the one used in the old auxiliary domain, or in other cases, the algorithm used in the old domain is even unknown to us. By the black-box adaptation property, we don't need to have any idea on the model used in the auxiliary do- main, but only the model predictions are required.
- Reducing the labelling cost: by adapting the auxiliary ranking model to the target domain, only a small number of samples need to be labelled,In Section 5, we'll experimentally demonstrate the proposed RA-SVM model is quite robust and well-performed, even with only a small number of training samples labeled.
- Reducing the computational cost: It has been shown that our ranking adaptation algorithm can be trans-formed into a Quadratic Programming(QP) problem, with the learning complexity directly related to the number of labeled samples in the target domain.

III. EXPLORE RANKING ADAPTABILITY

Though the ranking adaptation can mostly provide ben- efits for learning a new model, it can be argued that when the data from auxiliary and target domains share little common knowledge, the auxiliary ranking model can provide little help or even negative influence, to the ranking of the documents in the target domain. Consequently, it is imperative to develop a measure for quantitatively estimating the adaptability of the auxiliary model to the target domain. However, given a ranking model and a dataset collected for a particular target domain, it's nontrivial to measure their correlations di- rectly, because neither the distribution of the ranking model nor that of the labeled samples in the target domain is trivial to be estimated. Thus, we present some analysis on the properties of the auxiliary model,based on which the definition of the proposed adaptability is presented.

A.Auxiliary Model Analysis

We analyze the effects of auxiliary models through the loss constraint in the formulation of our RA-SVM. By substituting (4) into (2), we can obtain that:

$\sum f\ a\ (xijk\ ) + \Delta f\ (xijk\ ) \geq 1 - \xi ijk$
with $yij\ > yik\ ,$ and $\xi ijk\ \geq 0,$ (9)

where, as defined before, $xijk\ = \varphi(qi\ ,\ dij\ ) - \varphi(qi\ ,\ dik\ )$ and $\Delta f = \Sigma i,j,k\ \alpha ijk\ xT\ k\ x.$ Thus, in order to minimize the ranking error ξijk for the document pair dij and dik , we hope to get a large prediction value on the left-hand side of the first inequation in (9). For a given auxiliary ranking function f a, a comparatively large f a(xijk) suggests that f a can correctly judge the order for the document pair dij and dik , and vice versa. According to the constraints (9), if f a is capable of predicting the order of the documents correctly, we can correspondingly lower the contribution of the part of the ranking function learned in the target domain, i.e., Δf . At an extreme case, if f a is able to predict all pairs of documents correctly in the target domain, namely it can give perfect ranking lists for all the labeled queries, we may derive that f should be applied to the target domain directly with only small modifications, i.e., satisfying the "large margin"requirement in the target domain. On the other hand,if f cannot give a desirable ordering of the document pairs, we have to rely on Δf more to eliminate the side effects of f a , so that the ranking error over labelled samples is reduced. Consequently, the performance of f a over the labeled document pairs in the target domain can greatly boost the learning of RA-SVM for the ranking adaptation.

B.Ranking Adaptability

Based on the above analysis of f a , we develop the *ranking adaptability* measurement by investigating the correlation between two ranking lists of a labeled query in the target domain, i.e., the one predicted by f a and the ground-truth one labeled by human judges. Intuitively,if the two ranking lists have high positive correlation,the auxiliary ranking model f a is coincided with the distribution of the corresponding labeled data, therefore we can believe that it possesses high ranking adaptability towards the target domain, and vice versa. This is because the labeled queries are actually randomly sampled from the target domain for the model adaptation,and can reflect the distribution of the data in the target domain.

The proposed *ranking adaptability* measures the corelation between the ranking lists sorted by auxiliary model prediction and the ground truth, which in turn gives us an indication of whether the auxiliary ranking model can be adapted to the target domain, and how much assistance it can provide. Based on the *ranking adaptability*, we can perform automatic model selection for determining which auxiliary models will be adapted.

IV. PROCEDURE

This paper is integrated with following Modules:

A. Ranking Adaptation Module.
B. Explore Ranking adaptability Module.
C. Ranking adaptation with domain specific search Module.
D. Ranking Support Vector Machine Module.

A. Ranking adaptation Module

Ranking adaptation is closely related to classifier adaptation, which has shown its effectiveness for many learning problems. Ranking adaptation is comparatively more challenging. 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. We can adapt ranking models learned 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 and how to adapt the ranking model effectively and efficiently .

B.Explore Ranking adaptability Module

Ranking adaptability measurement by investigating the correlation between two ranking lists of a labeled query in the target domain, i.e., the one predicted by fa and the ground-truth one labeled by human judges. Intuitively, if the two ranking lists have high positive correlation, the auxiliary ranking model fa is coincided with the distribution of the corresponding labeled data, therefore we can believe that it possesses high ranking adaptability towards the target domain, and vice versa. This is because the labeled queries are actually randomly sampled from the target domain for the model adaptation, and can reflect the distribution of the data in the target domain.

C. Ranking adaptation with domain specific search Module

Data from different domains are also characterized by some domain-specific features, e.g., when we adopt the ranking model learned from the Web page search domain to the image 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. The basic idea of our method is to assume that documents with similar domain-specific features should be assigned with similar ranking predictions.

D. RankingSupportVector Machines Module

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 fa. Such a method is more advantageous than data based adaptation, because the training data from auxiliary domain may be missing or unavailable, for the copyright protection or privacy issue, but the ranking model is comparatively easier to obtain and access.

V. RELATED WORK

We present some works that closely relate to the concept of ranking model adaptation here. To create a ranking model that can rank the documents according to their achine learning techniques have been proposed. Some of them transform the ranking problem into a pairwise classification problem, which takes a pair of documents as a sample, with the binary label taken as the sign of the relevance difference between the two documents, e.g. Ranking SVM,RankBoost, RankNet and etc. Some other methods including ListNet,AdaRank,PermuRank,LambdaRank and etc., focus on the structure of ranking list and the direct optimization of the objective evaluation measures such as Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG). In this paper, instead of designing a new learning algorithm, we focus on the adaptation of ranking models across different domains based on the existing learning to rank algorithms.A lot of domain adaptation methods have also been proposed to adapt auxiliary data or classifiers to a new domain. Daume and Marcu proposed a statistical formulation in terms of a mixture model to address the domain distribution differences between training and testing set. A boosting framework was also presented for the similar problem . For natural language processing, Blitzer and et al introduced a structural correspondence learning method which can mine the correspondences of features from different domains. For multimedia application, Yang and et al.proposed Adaptive SVM algorithm for the cross-domain video concept detection problem. However, these works are mainly designed for

classification problems, while we focused on the domain adaptation problem for ranking in this paper.

VI. 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.

VII. CONCLUSION

As various vertical search engines emerge and the amount of verticals increases dramatically, a global ranking model, which is trained over a dataset sourced from multiple domains, cannot give a sound performance for each specific domain with special topicalities, document formats and domain-specific features. 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. 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, only the relevance predication of the auxiliary ranking models is needed for the adaptation.

Based on, two variations called margin rescaling slack rescaling are proposed to utilize the domain specific features to further facilitate the adaptation, by assuming that similar documents should have consistent rankings, and constraining the margin and loss of RA-SVM adaptively according to their similarities in the domain-specific feature space. Furthermore, we propose *ranking adaptability*, to quantitatively measure whether an auxiliary model can be adapted to a specific target domain and how much assistance it can provide

REFERENCES

[1] User Interfaces in C#: Windows Forms and Custom Controls by Matthew MacDonald.

[2] Applied Microsoft® .NET Framework Programming (Pro-Developer) by Jeffrey Richter.

[3] Practical .Net2 and C#2: Harness the Platform, the Language, and the Framework by Patrick Smacchia.

[4] Data Communications and Networking, by Behrouz A Forouzan.

[5] Computer Networking: A Top-Down Approach, by James F. Kurose.

[6] Gabriel R. Bitran and Rene Caldentey. An overview of pricing models for revenue management.N.Bruno and S.Chaudhuri. An online approach to physical design tuning.

[7] Xi-Ren Cao, Hong-Xia Shen, R. Milito, and P. Wirth. Internet Pricing with a game theoretical approach: concepts and examples.

[8] Ch Chen, Muthucumaru Maheswaran, and Michel Toulouse. Supporting co-allocation in an auctioning-based resource allocator for grid systems.

[9] S. Choenni, H. M.Blanken, and T. Chang. On the selection of secondary indices in relational databases.

[10] D. Dash, Y. Alagiannis, C. Maier, and A. Ailamaki. Caching all plans with one call to the optimizer.

AUTHORS

First Author – 1.M.Sirisha is pursuing M.Tech in Computer Science and Engineering from Nagole Institute of Technology and Science, JNTUH, A.P, INDIA. Her research areas include data mining, wireless mobile communication and network security

Second Author – SANJEEVA RAO SANKU ,Assoc. Prof. Dept. CSE, M.Tech,CSE(from JNTU Kakinada Campus,Completed in 2006)B.Tech(from Mahatma Gandi Institute of Technology, completed in 2003)Experience: 9+ years Interesting areas: Data warehousing and Data Minig, Computer Networks, Image Processing Attended number of workshops in various engineering colleges and universities. Conducted number of workshops in various colleges.number.