Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 4, April 2014, pg.103 – 109

REVIEW ARTICLE

Review on "Adaption of Ranking Model for Domain Specific Search"

Mr. Pratik R.Mantri[1], Prof. Mahip M.Bartere[2]

[1]ME (CSE), First Year, Department of CSE, G.H.Raisoni College of Engineering and Management, Amravati
Sant Gadgebaba Amravati University, Amravati, Maharashtra, India - 444727
pmpmyogesh@gmail.com

[2]Assitantant Professor, Department of CSE, G.H.Raisoni College of Engineering and Management, Amravati
Sant Gadgebaba Amravati University, Amravati, Maharashtra, India – 444727
mahip.bartere@raisoni.net

Abstract-Different new vertical domains are coming everyday so running a sophisticated ranking model Is no longer enviable as the domain are different and building a separate model for each domain is also not favorable because there much time required for labeling the data and training the samples. In this paper we are managing the above problem by regularization based algorithm called as ranking adaptation SVM (RA-SVM), the algorithm is used to adapt existing ranking model of sophisticated search engine to new domain. Here performance is still guaranteed and times taken to label the data training the samples are reduced. The algorithms only requires prediction from existing ranking model and do not require internal structure of it. Adapted ranking model concentrate on specific domain to achieve superior results which are relevant to the search, further it reduces the searching cost also as the most appropriate search results are shown.

Index Terms—Broad-based search; Regularization; Support vector machines; Adaption of model

I. Introduction

Now a day's people are more dependent on the internet for their day to day work, official work and academic work. As the result they want perfect result within short time. People perform their work Using search engines available on the net like Google, Bing, yahoo etc. they insert search query in it. Search is an operation where the inserted key words are sent over the networks to the information is shown to the user by the search engines [1]. The search engine uses broad based ranking model for retrieve the information and the ranking model of broad-based engine search is built upon the data from multiple domain. Search performed by User with specific search intention couldn't get the specific information as it fail to generalize the information. Focus is now moving broad based ranking model to domain specific search for performing the Special search intentions, gives the best results free from anomalies [2].Learning to rank is supervised learning technique where ranking model is to be learned through the repetitive machine learning process, Once the ranking model is learned it is hopefully capable of ranking the documents according to the query inserted by the user. Based on the machine learning technique there are many ranking algorithms e.g. lambda rank [3], rank net [4], List net [5], rank boost [6] etc. Further domain specific features can be used to further boost the search result, like content features of the image, videos or music.

II. Related Work

Adaptation of ranking model is closely related to classifier adaptation which was suffered from covariate shift and concept drifting. To create ranking model that can rank the document according to the significance to the given query some of them are classical BM25 [7] and language model for information retrieval [LMIR]. [8][9]They were best suited for the road based search engines with some of the parameters are needed to be familiar. On the other hand, with the expansion of algebraic knowledge method, and additional label information with compound features being accessible, sophisticated ranking models become more enviable for achieving better ranking performance. Newly, dozens of knowledge to rank algorithms base on machine learning system have been projected. A few of them convert the ranking problems into a paper-wise sorting problem, which take a couple of credentials as a model, with the binary labeled taken as the sign of the significance difference among two documents. A set of domain revision methods have also been projected to adapt supporting data or classifiers to a new domain. Daume and Marcu projected an arithmetic formulation in conditions of a combination model to address the field circulation difference between teachings and testing sets a boosting structure was also accessible intended for the analogous difficulty for expected language processing, Blitzer introduce. A structural association knowledge process which can extract the correspondence of facial appearance from unusual domains.

III. Ranking Adaption

We classify the ranking adaptation problem suitably as follow: designed for the target domain, a query position P = {p1, p2. . . pM } and a document locate L = {l1 , l2 , . . . , lN } are given. For each query pi ∈ P, a list of documents li = {li1 , li2 , . . . , li, n(pi) } are return and labeled with the significance degrees zi = {zi1 , zi2 , . . . , zi,n(pi) } by human annotators. The significance degree is generally a real value, i.e., zij ∈ R, so that unusual returned documents can be compared for sorting a prepared list. For each query document pair < pi, lij >, an s-dimensional query reliant feature vector φ (pi, lij) ∈ Rs is extracted, e.g., the term occurrence of the query keyword pi in the heading, body Of the paper lij. Some other hyperlink based fixed rank information is also considered, such as Page rank, HITS and so on. N (pi) denotes the quantity of return papers intended for uncertainty pi. The intention of learning to rank is to estimate a ranking function f ∈ Rs → R so that the documents l can be ranked for a given query p according to the value of the forecast f (φ (p, l)). In the setting of the projected level revision, equally the quantity of query M and the quantity of the return papers N (pi) in the instruction set are contained to be small. They are inadequate to learn a proficient ranking model for the target domain. However, a supporting ranking model fa, which is well skilled in another domain over the labeled data Pa and La, is accessible. It is implicit that the supporting ranking model fa contains a lot of preceding knowledge to rank documents, so it can be capable of to act as the bottom model to be modified to the innovative field. a small number of instruction samples can be enough to adapt the ranking model since the preceding knowledge is accessible. Before the introduction of our projected ranking adaptation algorithm, it's significant to evaluate the formulation of Ranking Support Vector Machines (Ranking SVM), that is mainly efficient learning to rank algorithms, and is here in work as the basis of our projected algorithm.

A. Ranking SVM

Analogous to the predictable Support Vector Machines (SVM) for the categorization problem, the inspiration of Ranking SVM is to find out a single dimensional of linear subspace, somewhere the point can be structured into the best possible level list in some criterion. Thus, the level function takes the structure of the linear representation f (φ (p, l)) = wT φ (p, l), where the preconception factor is unobserved, for the reason that the absolute ranking file sort by the guess f is invariant to the preconception. The optimization difficulty used for Ranking SVM is defined as follow:

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

(Here we use symbol q=p, d=l and y=z in above term)

Where C is the transaction constraint for matching the large-margin regularization ||f||2 and the failure expression □i, j, k ξijk since f is a linear representation, we can obtain that f (φ(pi , lij))−f (φ(pi ,lik)) = f (φ(pi , lij)−φ(pi , lik)), with φ(pi , lij)−φ(pi , lik) denoting the distinction of the feature vectors among the document pair lij and lik . If we further initiate the binary label sign (zij − zik) for each couple of documents lij and lik , the above Ranking SVM difficulty can be view as a usual SVM for classify paper pair into helpful or harmful, i.e., whether the paper lij should be ranked above lik or not. In view of the fact that

the quantity of labeled sample intended for the innovative domain is little, if we instruct the model use only the sample in the innovative domain, it will experience from the insufficient training sample difficulty, which is ill-posed and the outcome may be simply over specific to the label sample with low generalization facility. Additionally, the present SVM solver requires super-quadratic computational expenditure for the instruction; as impact, it is somewhat lengthy and nearly infeasible to train model using the instruction data from both the supporting domain and the objective domain. This difficulty is added severe for the ranking SVM since the instruction are based on pair and so the crisis amount is quadratic to the illustration size. Within the subsequent, we will extend and algorithm to label in the innovative domain. By model adaption, equally the efficiency of the result ranking model and the efficiency of the instruction procedure are achieved.

B. Ranking Adaptation SVM

It can be implied that, if the supporting domain and the objective domain are connected, their exacting ranking function fa and f must have analogous shapes in the utility space IRs! IR. Underneath such a proposition, f^a actually provides a preceding knowledge for the allocation of f in its constraint space. The conventional regularization construction, such as L_p-norm regularization, multiple regularization planned for SVM, regularized neural system, and so on, shows that the explanation of an ill-posed difficulty can be approximated from distinction standard, which contain mutually the data and the preceding statement. Consequently, we can adjust the regularization frame which develops the fast the former information, so that the ill-posed difficulty in the object domain, where only a small number of query paper pair is label, can be solve elegantly. By modeling our hypothesis into the regularization expression, the knowledge problem of Ranking Adaptation SVM can be formulate as

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

(Here we use symbol q=p, d=l and y=z in above term)

The intention function consists of the adaptation regularization expression $\|f - f^a\|^2$, which minimize the space between the objective ranking function and the supporting one in the utility space or the constraint space, to build them secure; the large-margin regularization $\|f\|^2$ and the failure expression $\Sigma_{i;j;k}\ \xi_{ijk}$. The parameter σξ (0, 1) is a tradeoff term to balance the contributions of large-margin regularization $\|f\|^2$which makes the learned model numerically steady, and adaptation regularization $\|f - f^a\|^2$ which makes the well-read model similar to the supporting one.

C. Optimization Methods

To optimize difficulty (2), we briefly indicate xijk = φ (pi, lij) − φ (pi, lik) and establish the Lagrange multipliers to combine the constraint of (2) into the intention function, which outcome in the primitive problem:

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

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

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

D. Discussions

The projected RA-SVM has a number of advantages, which makes our algorithm well relevant and flexible when applied to the practical application. We'll give more discussions of the uniqueness of RA-SVM in the following.

a) Model adaptation: the projected RA-SVM doesn't require the labeled instruction sample from the supporting domain, but simply its ranking model *f*a. Such a process is more beneficial than data-based adaptation, because the instruction data from supporting domain may be absent or unavailable, for the exclusive rights security or privacy issue, but the ranking model is relatively easier to realize and access.

b) Black-box adaptation: The internal demonstration of the model *f*a is not required, but only the Prediction of the supporting model to the instruction samples from the intention domain $f^a(x)$ is used. It brings a group of flexibilities in various situations where even the supporting model itself may possibly be unavailable. Also, in some cases, we would like to use an additional extremely developed algorithm for knowledge the ranking model for the innovative objective domain, than the one used in the previous supporting domain, or in further cases, the algorithm used in the previous domain is even unfamiliar to us. By the black-box adaptation property, we don't need to have any thought on the model used in the supporting domain, but only the model prediction are important.

c) Reducing the labeling cost: By adapting the supporting ranking model to the objective domain, only a little quantity of samples require to be labeled, while the lacking instruction sample problem will be address by the regularization term $\|f - f^a\|^2$, which really assigns a former to the objective ranking model.

d) Reducing the computational cost: It has been exposed that our ranking adaptation algorithm can be changed into a Quadratic Programming difficulty, with the learning complexity directly associated to the number of labeled samples in the objective domain. Platt projected the sequential minimal Optimization (SMO) algorithm which can partition a large QP problem into a sequence of sub problems and optimize them iteratively. The time complication is about O (n2.3) for general kernels. Cutting-plane technique is adopted to solve SVM for the linear kernel, which further reduces the time complication to O (n). Here, n is the number of labeled paper pairs in the objective domain. According to the above dialogue, the size of the labeled instruction set is deeply compact. Thus, n is considerably small, which in turn leads to the efficiency of our algorithm.

IV. RANKING ADAPTATION WITH DOMAIN-SPECIFIC FEATURE

Normally, data from unusual domains are also characterized by some domain-specific features, e.g., when we acknowledge the ranking model learned from the webpage search domain to the image search domain, the image substance can recommend extra information to make possible the text-based ranking model adaptation. In this section, we talk about how to consume these domain-specific features, which are typically composite to convert to textual representation directly, to further improve the performance of the proposed RA-SVM. The basic idea of our method is to assume that papers with analogous domain-specific features should be assigned with related ranking predictions. We name the above hypothesis as the consistency assumption, which implies that a strong textual ranking function should execute consequence prediction that is reliable to the domain-specific features. To implement the stability assumption, we are encouraged by the work and recall that for RA-SVM, the ranking failure is directly associated with the slack variable, which stands for the ranking failure for couple wise papers, and is nonzero as long as the ranking function predict an incorrect categorize for the two papers. In addition, as a large margin machine, the ranking failure of RA-SVM is also associated with the huge margin specified to the learned ranker. Consequently, to incorporate the uniformity constraint, we rescale the ranking failure based on two strategies, specifically margin rescaling and slack rescaling. The rescaling degree is controlled by the connection between the documents in the domain-specific quality space, so that analogous papers bring about less ranking failure if they are ranked in an incorrect order. We talk about in depth formulations of margin rescaling and slack rescaling as follow.

A. Margin Rescaling

Margin rescaling denotes that we rescale the margin violation adaptively according to their similarity in the domain-specific feature space. Particularly, the Ranking Adaptation SVM with Margin Rescaling (RA-SVM-MR) can be defined as the following optimization problem

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

Where $0 \leq \delta_{ijk} \leq$ ijk denote the similarity connecting document l_{ij} and l_{ik} return for query p_i in the domain-specific feature space. The above optimization problem differs from in the first linear inequity constraint, which varies the margin adaptively. Compared to a couple of unlike documents, similar ones with larger δ_{ijk} will result in a smaller margin to convince the linear constraint, which produce a lesser amount of ranking failure in conditions of a slighter slack variable ξ_{ijk} if the document pair l_{ij} and l_{ik} (namely p_{ijk}) is ranked in an incorrect order by the function f. The double problem is

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

And the desired ranking function takes the same form as, as shown above.

B. Slack Rescaling

Compared to margin rescaling, slack rescaling is projected to rescale the slack variables according to their similarity in the domain precise feature space. We define the analogous Ranking Adaptation SVM with Slack Rescaling (RA-SVM-SR) as the following optimization problem:

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

Distinct from margin rescaling, slack rescaling varies the amplitude of slack variables adaptively. If a couple of documents is dissimilar in the domain-specific feature space, by dividing 1 $\sigma_{ijk,}$ the slack variables that manage the ranking failure of the two papers are in the same way improved in categorize to satisfy the first linear parity, and vice versa. The double problem of is and the explanation of the standing function, as for RA-SVM-MR, is identical to, as shown in. It can be observed from the double format of that, slack rescaling is equal to rescaling the transaction parameters C for each couple wise documents, based on their similarity.

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

The optimizations of RA-SVM-MR and RA-SVM- SR have the exactly same time complication as for the RA-SVM, i.e., O $(n^{2:3})$ by using SMO algorithm and OðnÞ by means of cutting plane algorithm for the linear kernel.

TABLE 1
Ranking Adaptation Data Set Information

Dataset	#Query	#Query-Document	Relevance Degree	Feature Dimension
TD2003	50	49171	2	44
TD2004	75	74170	2	44
Web Page Search	2625	122815	5	354
Image Search	1491	71246	3	354

Therefore, although domain-specific features are Integrated for the model adaptation, we didn't get scheduled any additional efficiency problems.

V. CONCLUSION

As a variety of upright search engines appear and the quantity of verticals increases significantly, a global ranking model, which is qualified over a dataset sourced from several domains, cannot give a sound performance for each exact domain with particular topicalities, document formats and domain-specific features. Structure one model for every vertical domain is both difficult for labeling the data and time-consuming for learning the model. In this paper, we suggest the ranking model adaptation, to adapt the well learned models from the sophisticated search or any other supporting domains to a new target domain. By model adaptation, only a little number of samples requires to be labeled, and the computational cost for the training process is greatly compact. Based on the regularization framework, the Ranking Adaptation SVM algorithm is projected, which performs adaptation in a black-box way, only the significance predication of the supporting ranking models is desired for the adaptation. Based on, two variations called margin rescaling slack rescaling are projected to consume the domain specific features to further facilitate the adaptation, by assuming that similar documents should have steady rankings, and constraining the margin and loss of RA-SVM adaptively according to their similarities in the domain-specific-feature-space.

a) The predictable RA-SVM can enhanced develop equally the supporting models and intention domain labeled queries to learn a stronger ranking model for the intention domain information.

b) The consumption of domain-specific features can steadily further develop the model adaptation, and RA-SVM-SR is somewhat more strong than RA- SVM-MR.

c) The adaptability level is consistent to the convenience of the supporting model, and it can be deemed as a proficient criterion for the supporting model selection.

d) The projected RA-SVM is as proficient as directly knowledge a model in an objective domain, while the incorporation of domain-specific features does not bring much learning obstacle for algorithms RA- SVM-SR and RA-SVM-MR.

References

[1] M.Belkin, P. Niyogi, and V. Sindhwani, "Manifold Regular- ization: A Geometric Framework for Learning from Labeled and Unlabeled Examples," J. Machine Learning Research, vol. 7, pp. 2399-2434, Nov. 2007.

[2] J. Blitzer, R. Mcdonald, and F. Pereira, "Domain Adaptation with Structural Correspondence Learning," Proc. Conf. Empirical Meth- ods in Natural Language Processing (EMNLP '06), pp. 120-128, July 2006.

[3] 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.

[4] C.J.C.Burges,T.Shaked,E.Renshaw,A.Lazier,M.De-eds,N.Hamilton , G.Hullender , "Learning to Rank Using Gradient Decent ",Proc.22th Int'l Conf. machine Learning (ICMC'05),2005.

[5] Z.Cao and T.Yan Liu,"Learning to Rank: From pair wise Approach to lead wise Approach,"proc.24th Int'l conf.Machine Learning (ICML 07), pp.129-136, 2007.

[6] Y.Freund,R.Iyer ,R.E.Schapire,Y Singer , and G.Ditterich ,"An Efficient Boosting Algorithm for combining Preference,"J.Machine Learning Research vol.4,pp.933-969,2003.vol.30 nos.1/2,pp
81-93.june 2007.

[7] H.Shimodaria,"Improving Predictive Inference Under covariate shift by Weighting the log-Likelihood function,"J.Stastical Planning and Inference vol.90, no.18.227-244, 2000.

[8] J.Lafferty and C.Zhai,"Document Language Models, Query Model, and Risk Minimization for Information Retrieval,"Proc.24th Ann.Int'l ACM SIGIR cont.Research and Development in Information Retrieval (SIGIR 01),

pp.111-119, 2001.

[9] 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, 2004.

[10] L .Page, S .Brin, R .Motwani, and T.Winogard. "The Pagerank Citation Ranking: Bringing order to the web, "Technical report, Stanford Univ., 2005.

[11] J.M.Kleinberg, S.R.kumar, P.Raghvan, S.Rajgopalan, And a Tomkins, "The web as a Graph: Measurement, Model and Method,"proc.Int'l Conf Combinatorial and Computing, pp.1-18, 2008.

[12] V.N.Vapnik Statistical Learning Theory, Wiley Interscience, 2004.

