ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 2, February 2013

RANKING ADAPTATION SVM FOR TARGET DOMAIN SEARCH

M.S.Gayathri[1], S.Leela[2], Asst. Prof
Department of Computer Science and Engineering, Karunya University
Coimbatore, India

***Abstract*: With the growth of different search engines, it becomes difficult for an user to search particular information effectively. If a search engine could provide domain specific information such as that confines only to a particular topicality, it is referred to as domain specific engine. Applying the ranking model trained for broad-based search to a domain specific search does not achieve good performance because of domain differences. Building a different ranking model for each domain is laborious and time consuming. In this paper, the ranking model used in broad-based search is adapted to domain specific search. An algorithm called Ranking Adaptation SVM is used to effectively adapt a ranking model to a target domain. Such an adaptation has advantages since it needs only the predictions from existing ranking models. Ranking Adaptability measurement is used to quantitatively estimate whether a ranking model can be adapted to new domain.**

***Keywords*: Domain Specific engine, Support vector machine, information retrieval.**

I.INTRODUCTION

Machine learning [1] is an application of a learning process that is used extensively. This is a vast field that includes learning from examples and reinforcement learning (learning from teacher etc).

A learning algorithm takes a data and its surrounding information as input and returns a statement as output. Machine learning analyses previous happenings and their outcomes and learns how to use these and make generalisations about new cases.

A machine learning system uses lot of information about the environment which makes an entire finite set called the training set. This training set contains examples i.e. observations. The training set is finite hence not all concepts can be learned exactly.

The ranking involves learning a ranking model from a training set such that it produces a "good" ordering on the test set.

Information retrieval [12], [19] is the method of obtaining information relevant to the need of an user from a collection of information resources.

Learning to rank [6], [9] is a supervised or semi-supervised machine learning problem that constructs a ranking model from training data. This ranking order is typically induced by giving a numerical or ordinal score or a binary judgment (e.g. "relevant" or "not relevant") for each item.

Domain-specific Web search engines are effective tools as cited by (Joachims .T et al) for reducing the difficulty experienced when acquiring information from the Web. A domain-specific search engine can be made simply by adding domain-specific keywords, to the user's input query and forwarding it to a general-purpose Web search engine. Keywords can be effectively discovered from Web documents using machine learning technologies.

In machine learning [1], support vector machines (SVMs, support vector networks also) are one of the supervised learning models with associated learning algorithms that analyze data. It is used for classification and regression analysis. The centre of Support Vector Machine is that it uses a set of input data and predicts, for each given input, a relevant output is produced. An SVM [16] model is a representation of the examples such as few points in some space, related so that the examples of the separate domains are divided by a clear gap that is as wide as possible.

In this paper, the adaptation of ranking models is focussed 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 ADAPTATION

Ranking adaptation can be formally defined for the target domain as follows: a query set Q= $\{q_1,q_2.....q_m\}$ and a document set D=$\{d_1,d_2....d_n\}$ are given. For each query $q_i \in Q$, a list of documents $d_i=\{d_{i1},d_{i2} ....d_{i,n(qi)}\}$ returned and labeled with the relevance degrees $Y_i=\{y_{i1},y_{i2},.....y_{i,n(qi)}\}$ by human judges. The relevance degree is generally a real value, i.e., $y_{ij} \in IR$, so that different returned documents can be compared for sorting an ordered list. For each query document pair $<q_i ,d_{ij}>$, an s-dimensional query dependent feature vector $Ó(q_i, d_{ij}) \in IR^s$ is obtained, e.g., the query keyword's term frequency q_i in the title, body, URL of the document d_{ij}. $n(q_i)$ denotes the number of returned documents for query q_i. The objective of the learning to rank is to calculate approximately a ranking function $f \in IR^s \rightarrow IR$ so that the documents d can be ranked for a given query q according to the value of the prediction f(Ó(q,d)).

In the proposed ranking adaptation [, both the query numbers m and the number of the retrieved documents $n(q_{i)}$ in the training set are assumed to be small. So this becomes insufficient to learn the ranking model for target domain. But, an auxiliary ranking model f^a, which is much trained in another domain other than the target domain over the labelled data Q^a and D^a, is available. It is understood that the auxiliary ranking model f^a contains a lot of old knowledge to rank documents, so this is used as a base model to be used in target domain. Since the prior knowledge is available only few training samples are needed to adapt the ranking model . Before describing the adapted Ranking model that is the Ranking Adaptation SVM formulation, it is necessary to review the working of Ranking SVM which is the base of the method that is used.

III RANKING SVM

Similar to the basic SVM, the motivation of Ranking SVM [7] 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 $f(Ó(q,d))=W^TÓ(q,d)$, 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.

IV RANKING ADAPTATION SVM

It is supposed that if the target domain and the auxiliary domain are to be related then their respective ranking function f and f^a have similar shapes in the function space IRs−>IR. Under this statement, f^a really provides a past knowledge for the distribution of f in their parameter space. Some regularization framework like L_p-norm regularization, manifold regularization that are designed for SVM shows that the problem raised can be compensated using variational principle. Therefore, it is that they can adapt the regularization framework that utilizes the ranking function of the target domain f^a as the past information, so that the problem in the target domain, where only very less documents are labeled, can be solved sophisticatedly. The Ranking Adaptation SVM's [4] learning problem can be formulated as,

$$\min_{f,\pounds ijk} \frac{1-\partial}{2} ||f||^2+\partial/2\ ||f-f^a||^2 +C\sum_{i,j,k} \pounds_{ijk} \quad (1)$$

The adaptation regularization term in the objective function is $||f-f^a||^2$. This regularization term minimizes the distance between the ranking functions. i.e., the ranking function in auxiliary domain and the target domain of function space or the parameter space, to make them close. The parameter $\partial \in [0,1]$ is a trade off term. This is to balance the contributions of large-margin regularization $||f||^2$ and adaptation regularization $||f-f^a||^2$. Large-margin regularization term makes the learned ranking model to be numerically stable. When ∂=0, RA-SVM is equal to directly learn Ranking SVM over the target domain, without the adaptation of the ranking function of the auxiliary domain f^a.

V RANKING ADAPTABILITY

Based on the analysis of the ranking function f^a, develop the ranking adaptability measurement by studying the correlation between two lists from the ranking of a labeled query in the target or new domain, i.e., the one predicted by f^a and the other one labeled by human annotators. Naturally if both the two lists of ranking have high positive correlation, the auxiliary ranking model with ranking function f^a coincides with the distribution of the corresponding labeled data.

Therefore we can consider that it possesses high-ranking adaptability towards the target domain and the target domain will be possessed with characteristics of high ranking features. It is that the labeled queries are in fact arbitrarily sampled from the target domain for the adaptation of ranking model, and this can reflect the distribution of the data in the target domain. Here, we implement the known Kendall's equation to calculate the correlation between the two lists of ranking, and via this equation the proposed ranking adaptability is defined. Therefore, a more general definition for the correlation is,

$$\tau_i(f)=(N_i^c-N_i^d) / (N_i^c+N_i^d) \quad (2)$$

VI ADAPTATION FROM MULTIPLE DOMAINS

RA-SVM is extended to a more general setting, where models to rank a page learned from multiple domains are given. The RA-SVM for the multiple domain adaptation can be formulated as given below,

$$\min_{f,\xi_{ijk}} \frac{1-\partial}{2} \|f\|^2+\partial/2\sum_{r=1}^{R} \emptyset\|f-f^a\|^2 +C\sum_{i,j,k} \xi_{ijk} \quad (3)$$

VII RANKING ADAPTATION WITH DOMAIN-SPECIFIC FEATURE

Data from different domains are characterized by certain domain specific features, e.g., When we adopt the ranking model that is used in a webpage search domain to an image search domain [10], the information surrounding the image can also provide additional information to support text based ranking model. Here the domain specific features are utilized. These domain specific features are difficult to translate into textual forms. This boosts the performance of RA-SVM. The rule is that documents with similar domain specific features have to be ranked with similar rankings in the domain. This assumption is called as consistency assumption.

In order to implement the consistency assumption, the loss in ranking is directly related with the slack variable that stands for pair wise documents, and is kept nonzero as long as the ranking function used predicts a wrong order for the two documents. Therefore, to include the consistency constraint, we rescale the ranking loss based on two strategies, called margin rescaling [4] and slack rescaling (Bo Geng et al). The rescaling degree is controlled by the similarities between the documents in the domain-specific feature space, so that same documents bring about less ranking loss if they are ranked in a wrong order.

The architecture of RA-SVM is mentioned in [5] as given below,

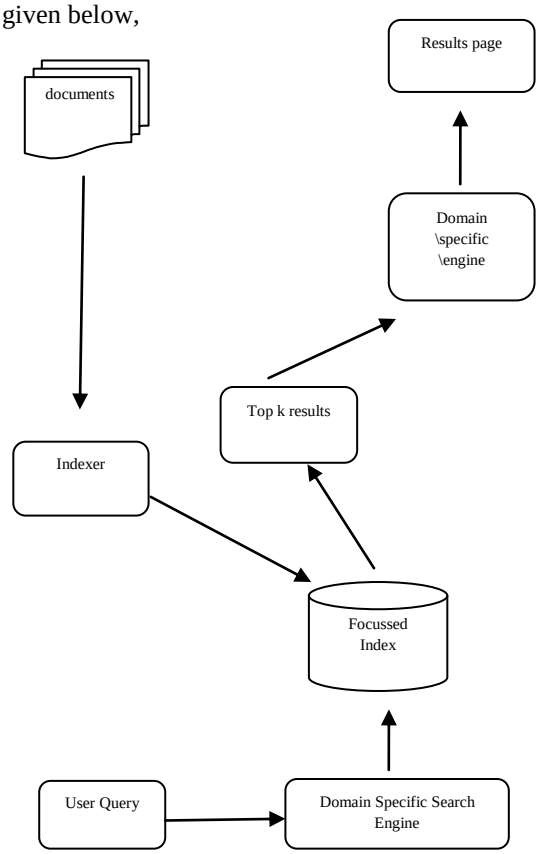


Fig 1. RA-SVM architecture

VIII RA-SVM MODULES

The RA-SVM system has been identified to have the following modules:

- Adaptation of Ranking Module.
- Explore Adaptability of Ranking Module.
- Adaptation of Ranking model with domain specific search Module.
- Ranking with Support Vector Machine Module.

a. Adaptation of Ranking Module.

Ranking adaptation is closely related to adaptation using classifiers, which has shown its effectiveness for many learning problems. Ranking adaptation is quite more challenging. Contrast to classifier adaptation, which mainly deals with binary targets, ranking adaptation adapts the model which is used to predict the rankings for a collection

of domains. It involves of prediction of domain to which adaptation can be done. In ranking the relevance between different domains it is that they are sometimes different and need to be aligned. Ranking models learned for the existing general search can be used to be adapted in 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 Adaptability of Ranking Module

Ranking adaptability *is* measured 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 it is believed that it possesses high ranking adaptability towards the target domain. It is because the labeled queries are actually randomly sampled from the target domain for the model adaptation, and can reproduce in the distribution of the data in the target domain.

c. Adaptation of Ranking Model With Domain Specific Search Module

Data from different domains are also characterized by some domain-specific features, e.g by adopting the ranking model learned from the Web page search domain to the image search domain, the content in the image can provide additional information to facilitate the text based ranking model adaptation. In this section, how to utilize these domain-specific features are examined, which are usually difficult to translate to textual forms directly, to further enhance the performance of the RA-SVM (Bo Geng et al 2012). The basic idea of the proposed method is to assume that documents with same domain-specific features should be given with similar ranking predictions. The above assumption is said as the consistency assumption, implies that a robust ranking function performs relevance prediction that is constant to the domain-specific features.

d. Ranking with Support Vector Machines Module

Ranking Support Vector Machines (Ranking SVM), which is one among the most effective learning to rank methods, is here used as the basis of the method used here. The RA-SVM does not need the labeled training samples from the auxiliary domain, but only its ranking model f^a. This method is more advantageous than data based adaptation, because the data that is trained from auxiliary domain may be missing or unavailable, due to copyright protection or privacy issue, but the ranking model used to rank a page is comparatively easier to obtain and access.

IX EXPERIMENTS

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

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. It has details pertaining to say for example two domains. i.e medical and product. Medical domain has details regarding certain specific medical information. Product domain has details pertaining to certain details about products like laptop etc.

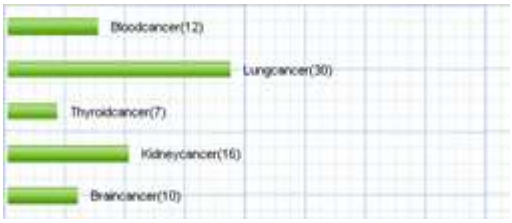
Search page:

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.

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 medical domain is created. The user is allowed to search for a keyword “medical”. Then under medical when cancer is searched for the most viewed pages is ranked first using RA-SVM. The following graph shows that the lung cancer is viewed the maximum times and is ranked higher than any other pages.



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 labelling 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

In the adaptation of the ranking model only their model prediction is needed. The internal representation of the auxiliary ranking model is not taken into consideration.

Reducing the labeling cost

Only a very few number of samples are to be labeled while adapting the auxiliary ranking model. Hence, the labelling 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.

X RELATED WORKS

There are some other related works closely related to the concepts of ranking models. The one used as the basis of the paper is Ranking SVM, that takes the form of RA-SVM. Some of the boosting ranking models for ranking a page is RankBoost [10].

A pair wise neural ranking algorithm called RankNet [7], [16] ranks pages effectively and is used extensively in the real world. The heart of the Google is the ranking model called the PageRank [17].

Inorder to reduce the fidelity loss a ranking method is posed called FRank [18]. Another ranking model called AdaRank [14] is efficient. LETOR [20], a benchmark dataset is used for several experimental verifications.

XI CONCLUSION

Thus, the adapted Ranking SVM performs better than other ranking models since it has lot of advantages. Model adaptation, reduced labelling cost, black box adaptation are the merits. In this method only the model is adapted, the internal representation is not taken into consideration. Ranking adaptation SVM works well with domain specific search engine.

ACKNOWLEDMENT

The authors would like to thank all the professors and lecturers who helped in completing this paper as well we thank the reviewers for their valuable comments and suggestions.

REFERENCES

[1] Andrew McCallum, Kamal Nigam, Jason Rennie, Kristie Seymore, (2004) ”A Machine Learning Approach to Building Domain-Specific Search Engines”.

[2] 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.

[3] Bo Geng, Linjun Yang, Chao Xu, and Xian-Sheng Hua, (April 2012) “Ranking Model Adaptation for Domain-Specific Search”, IEEE Transaction on Knowledge and Data Engineering, vol 24, No. 4.

[4] Bo Geng, Linjun Yang, Chao Xu, Xian-Sheng Hua, (Apr 26,2012) ”Ranking Model Adaptation for Domain-Specific Search”, United States Patent Publications.

[5] Burges. C.J.C et al Shaked. T, Renshaw. E, Lazier, A,Deeds. M, Hamilton. B, and Hullender. G, (2005), “Learning to Rank Using Gradient Descent”, Proc. 22th Int’l Conf. Machine Learning (ICML ’05).

[6] Chapelle. O, Keerthi. S, (July 20, 2009) ”Efficient Algorithms for Ranking with SVMs”, Information Retrieval Journal.

[7] Christopher J.C. Burges, ”From RankNet to LambdaRank to LambdaMART: An Overview” , Microsoft Research Technical Report MSR-TR-2010-82.

[8] Christopher J.C. Burges, Robert Ragno, Quoc Viet Le, (2006) “Learning to Rank with Nonsmooth Cost Functions”, Proc. Advances in Neural Information Processing Systems (NIPS ’06), pp. 193-200.

[9] Cui. C, Wen. F, and Tang. X, (2008) “Real Time Google and Live Image Search Re-Ranking”,

Proc. 16th ACM Int'l Conf. Multimedia, pp. 729-732.

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

[11] 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. 197-206.

[12] Jarvelin. K and Kekalainen. K, (2000) "IR Evaluation Methods for Retrieving Highly Relevant Documents", Proc. 23rd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '00), pp. 41-48.

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

[14] Jun Xu, Hang Li,(2007), "AdaRank: A Boosting Algorithm for Information Retrieval", Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 391-398.

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

[16] Krysta M. Svore, Lucy Vanderwende, Christopher J.C. Burges, " Enhancing Single-document Summarization by Combining RankNet and Third-party Sources", Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language.

[17] Matteo Pasquinelli, "Google's PageRank Algorithm: A Diagram of the Cognitive Capitalism and the Rentier of the Common Intellect".

[18] Ming-Feng Tsai, Tie-Yan Liu, (2000) "FRank: A Ranking Method with Fidelity Loss", SIGIR'07, July 23–27, Amsterdam, The Netherlands

[19] Ponte. J. M and Croft. W. B, (1998)"A Language Modeling Approach to Information Retrieval", Proc. 21st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 275-281.

[20] Tao Qin ,Tie-Yan Liu ,Jun Xu , Hang Li, (2007), "LETOR: A Benchmark Collection for Research on Learning to Rank for Information Retrieval", Proc. SIGIR Workshop Learning to Rank for Information Retrieval (LR4IR '07).