International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 1 Issue 1 –AUGUST 2013.

Domain based Ranking Adaptation Model in Search Engines

Bhu Devi Guggilam[#1], Damarla Sree latha[*2]

[#]*PG Scholar, Dept. of CSE, VRS & YRN College of Engg. & Technology, Chirala, AP*

[1] devigood1@gmail.com

[*]*Assistant Professor, Dept. of CSE, VRS & YRN College of Engg. & Technology, Chirala, AP*

***Abstract*—In recent years, Domain-Specific search engines are becoming popularly of their accuracy and it contains extra features on the web search engines, in the case of general extra benefits are not applicable. Though, it provides many services to the users, it contains some issues on the time consuming. The time consuming is to maintain is very difficult in this services. In the emerging search engines domains, to implement a broad-based ranking model directly to different domain is no longer desirable due to the domain differences, while building a unique ranking model for each domain is both laborious for labelling data and time consuming for training models. In this paper, we are proposing the new algorithm of ranking model adaption for domain-specific search to overcome the difficulties in the existing ranking model in the SVM. The new proposing algorithm is based on the regularization and to promote the old ranking model into the new domain. The training cost and the amount for the labelled data is to be reduced while the performance is still guaranteed. Our proposing algorithm requires only the augury from the existing models, rather than their internal representations or the data from assistant domains. We assume that the documents similar in the domain-specific feature space should have consistent rankings. Extensive analytical and experimental results are presented which show the scalability and efficiency of our proposing RA-SVM.**

***Index Terms*— Information retrieval, support vector machines, learning to rank, domain adaptation, ranking adaption model, SVM.**

I. INTRODUCTION

A search engine is a program [33] that can search the Web on a specific topic. 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 labelled with their relevancies to some queries. Then the model is hopefully capable of ranking the documents returned to new queries groups automatically. The performances of the learning to rank algorithm have already showed in the information retrieval. It is especially in the web search services.

In recent years, Domain-Specific search engines are becoming popularly of their accuracy and it contains extra features on the web search engines, in the case of general extra benefits are not applicable. . In the emerging search engines domains, to implement a broad-based ranking model directly to different domain is no longer desirable due to the domain differences, while building a unique ranking model for each domain is both laborious for labelling data and time consuming for training models. There are many vertical search engines are introduced in recent years, they contains different domain-specific features, document types and topicalities. For example, a image search engine deal with specialized in terms of its topical focus only. Whereas many music or videos search engine would concern only the documents in the particular formats. The techniques are broad-based and vertical search engines are most popular for the text search formats in recently. In the ranking model learned broad based can be utilized directly to rank the documents for the verticals.

For example web images are considered as text-based documents that is share with similar ranking features as a document or web page ranking, and text –based ranking model can be applied here directly. Nevertheless, the broad based ranking models are built upon the data from the multiple domains; it would not be generalize for particular domain. The broad based ranking model can only utilize the vertical domain's ranking features that are same to the broad-based domains for ranking, while the domain-specific features, such as the content features of images, videos, or music cannot be utilized directly. And those features are generally important for the semantic representation of the documents. It should be utilized to build a more robust ranking model for the particular vertical. From this, the each vertical learn its

own ranking model independently. Though, it provides many services to the users, it contains some issues on the time consuming. The time consuming is to maintain is very difficult in this services.

As the result of the ranking model of the broad based search can provides a number of reasonable, but not as perfect as specifically trained for the vertical search applications. Therefore we analyse the complexity in the broad based model and it is tradeoffs to the independent learning of a new ranking model for the vertical. the existing broad-based ranking model provides a lot of common information in ranking documents, only few training samples are needed to be labelled in the new domain. So that, the broad based ranking model can be completely adapted to the new domains with many specific features. And then based on the view of the broad based ranking model provides a prior knowledge and it consist of number of labelled samples are sufficient of the target domain ranking models. Hence, the costs for new verticals search are reduced, and auxiliary ranking models to the new target domain and to make full use of their domain-specific features.

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 labelled 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 twofold: 1) in ranking, the mainly concerned is about the preference of two documents or the ranking of a collection of documents, which is difficult to be modelled by classification or regression; 2) the relevance levels between different domains are sometimes different and need to be aligned.

In this paper, we mainly focus on the issues of an existing model. In order to overcome the issues new technique in proposing in this paper namely Ranking Model Adaptation for the domain-specific search (RA-SVM). In the ranking models, instead of utilizing the labelled data from auxiliary domains directly, it should be inaccessible due to the privacy issues or data missing. Model adaptation is more desirable than data adaptation, because the learning complexity is now only correlated with the size of the target domain training set, which should be much smaller than the size of auxiliary data set.

In this paper, we investigate the problems of an existing model:

1. Whether the amount of labelled data in the target domain is reduced while the performance requirement is still guaranteed?
2. How to adapt the ranking model effectively and efficiently?
3. How to utilize domain-specific features to further boost the model adaptation?

So that, in this paper we address the above three issues. To solve the first problem the proposed ranking adaptability measure, this quantitatively estimates and predicts to the potential performance for the adaptation. And to solve the second issue we proposed the algorithm as ranking adaptation SVM (RA-SVM). Our algorithm is a black box ranking model adaptation. The black-box adaptation property not only achieved flexibility and also the efficiency. To resolve the third problem, we assume that documents similar in their domain-specific feature space should have consistent rankings, e.g., images that are similar in their visual feature space should be ranked into similar positions and vice versa. We implement this idea by constraining the margin and slack variables of RA-SVM adaptively, so that similar documents are assigned with less ranking loss if they are ranked in a wrong order.

The rest of the paper is organized as follows. In Section II, we discuss about the related work of the domain-specific search models. In Section III formally introduces our proposed system of the paper. In Section IV we summarize about the algorithm used in the model. In Section V, we present the full simulation study of the proposed scheme. Finally, we conclude the paper and discuss future work in Section VI.

II. Related Works

In this section, we briefly discuss the works which is similar techniques as our approach but serve for different purposes.

Andrew McCallum, Kamal Nigam, Jason Rennie and Kristie Seymore [34], this paper proposes the use of machine learning techniques to greatly automate the creation and maintenance of domain-specific search engines. We describe new research in reinforcement learning, text classification and information extraction that enables efficient spidering, populates topic hierarchies, and identifies informative text segments. Using these techniques, we have built a demonstration system: a search engine for computer science research papers available at www.cora.justresearch.com.

Shilpy Sharma [33], this paper describes the use of semi-structured machine learning approach with Active learning for the “Domain Specific Search Engines”. A domain-specific

search engine is "An information access system that allows access to all the information on the web that is relevant to a particular domain. The proposed work shows that with the help of this approach relevant data can be extracted with the minimum queries fired by the user. It requires small number of labeled data and pool of unlabelled data on which the learning algorithm is applied to extract the required data.

Sukanta Sinha, Rana Dattagupta, and Debajyoti Mukhopadhyay [35], Search Engine is a Web-page retrieval tool. Nowadays Web searchers utilize their time using an efficient search engine. To improve the performance of the search engine, we are introducing a unique mechanism which will give Web searchers more prominent search results. In this paper, we are going to discuss a domain specific Web search prototype which will generate the predicted Web-page list for user given search string using Boolean bit mask.

Satoshi Oyama, Takashi Kokubo, and Toru Ishida [36], — Domain-specific Web search engines are effective tools for reducing the difficulty experienced when acquiring information from the Web. Existing methods for building domain-specific Web search engines require human expertise or specific facilities. However, we can build a domain-specific search engine simply by adding domain-specific keywords, called "keyword spices," to the user's input query and forwarding it to a general-purpose Web search engine. Keyword spices can be effectively discovered from Web documents using machine learning technologies. This paper will describe domain-specific Web search engines that use keyword spices for locating recipes, restaurants, and used cars.

Chunxia YIN, Jian LIU, Chao YANG, and Huiying ZHANG[37], A focused crawler is a Web crawler aiming to search and retrieve Web pages from the World Wide Web, which are related to a domain-specific topic. Rather than downloading all accessible Web pages, a focused crawler analyzes the frontier of the crawled region to visit only the portion of the Web that contains relevant Web pages, and at the same time, try to skip irrelevant regions. In this paper, we present a new crawling strategy that employed the C4.5 decision tree to predict the relevance of a link target, and combined the algorithm with the link characteristic of parent pages. Experimental results indicate that the new crawling method has better performance, and it was able to fetch higher topic relevant information.

Sofia Ceppi [38], My Ph.D. thesis focuses on the design of economic mechanisms inspired by sponsored search auctions to support new generation search engines. These engines (called integrators) are based on multi-domain queries and on the federation of multiple domain-specific search engines. The problem studied in my thesis is essentially a mechanism design problem where two levels are present: in the first, the advertisers submit bids to the domain-specific search engines; in the second, the domain-specific search engines interact with an integrator in the attempt to produce the best search results.

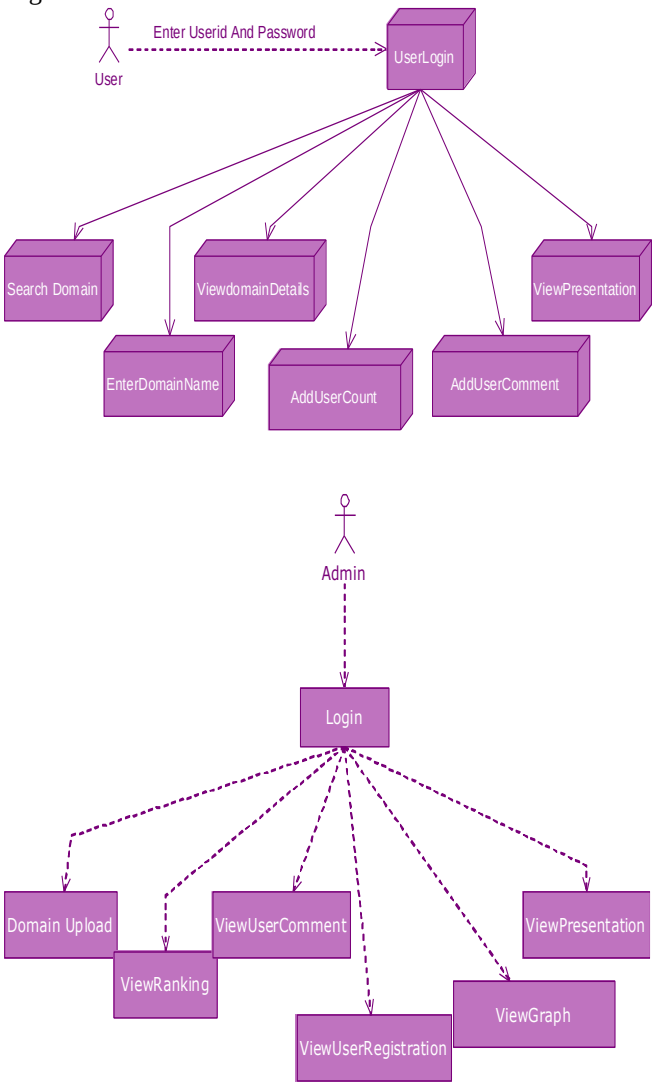
We present some works that closely relate to the concept of ranking model adaptation here. To create a ranking model that can rank the documents according to their relevance to a given query, various types of models have been proposed, some of which have even been successfully applied to web search engines. Classical BM25 [24] and Language Models for Information Retrieval (LMIR) [19], [23] work quite stable for the broad-based search with few parameters needing adjusted. However, with the development of statistical learning methods, and more labeled data with complicated features being available, sophisticated ranking models become more desirable for achieving better ranking performance. Recently, a dozen of learning to rank algorithms based on machine 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 [12], [14], RankBoost [9], RankNet [4], and, etc. Some other methods including ListNet [5], SVMMap [31], AdaRank [28], PermuRank [29], LambdaRank [3], and, etc., focus on the structure of ranking list and the direct optimization of the objective evaluation measures such as Mean Average Precision and Normalized Discounted Cumulative Gain. 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. For natural language processing, Blitzer et al. [2] introduced a structural correspondence learning method which can mine the correspondences of features from different domains. For multimedia application, Yang et al. [30] 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. Ranking adaptation is closely related to classifier adaptation, which has shown its effectiveness for many learning problems [2], [7], [8], [25], [18], [30], [32].

III. PROPOSED WORK

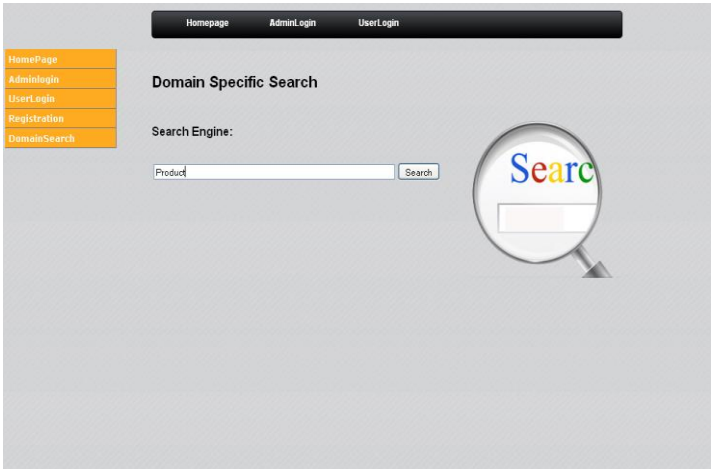
In this paper, we mainly focus on the issues of an existing model. In order to overcome the issues new technique in proposing in this paper namely Ranking Model Adaptation for the domain-specific search (RA-SVM). In the ranking models, instead of utilizing the labelled data from auxiliary domains directly, it should be inaccessible due to the privacy issues or data missing. Model adaptation is more desirable than data adaptation, because the learning complexity is now only correlated with the size of the target domain training set,

which should be much smaller than the size of auxiliary data set.

There are three issues raised in an existing ranking model. To solve the first problem the proposed ranking adaptability measure, this quantitatively estimates and predicts to the potential performance for the adaptation. And to solve the second issue we proposed the algorithm as ranking adaptation SVM (RA-SVM). Our algorithm is a black box ranking model adaptation. The black-box adaptation property not only achieved flexibility and also the efficiency. To resolve the third problem, we assume that documents similar in their domain-specific feature space should have consistent rankings, e.g., images that are similar in their visual feature space should be ranked into similar positions and vice versa. We implement this idea by constraining the margin and slack variables of RA-SVM adaptively, so that similar documents are assigned with less ranking loss if they are ranked in a wrong order.



IV. Simulation Works/Results



V. Conclusion

In this paper, we proposed the ranking model adaptation, and it adapts the well learned models from the broad based search into a new target domain. By newly adapted model consists of small number of samples need to be labeled, and the computational cost for the training process is greatly reduced. In this Ranking Adaptation SVM algorithm is proposed, it is based on the regularization framework and it performs adaptation in a black-box way. Based on RA-SVM, two variations called RA-SVM margin rescaling and RA-SVM 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. There are many experiments is carried on to improve the performance of the proposed ranking model. Based on the results we can derive the following as conclusions:

The proposed RA-SVM can better utilize both the auxiliary models and target domain labeled queries to learn a more robust ranking model for the target domain data. The utilization of domain-specific features can steadily further boost the model adaptation and RA-SVM-SR is comparatively more robust than RASVM- MR. The adaptability measurement is consistent to the utility of the auxiliary model, and it can be deemed as an effective criterion for the auxiliary model selection. The proposed RA-SVM is as efficient as directly learning a model in a target domain, while the incorporation of domain- specific features doesn't brings much learning complexity for algorithms RASVM- SR and RA-SVM-MR.

VI. REFERENCES

[1] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples," J. Machine Learning Research, vol. 7, pp. 2399-2434, Nov. 2006.

[2] J. Blitzer, R. Mcdonald, and F. Pereira, "Domain Adaptation with Structural Correspondence Learning," Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP '06), pp. 120-128, July2006.

[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. Deeds, N. Hamilton, and G. Hullender, "Learning to Rank Using Gradient Descent," Proc. 22th Int'l Conf. Machine Learning (ICML '05), 2005.

[5] 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.

[6] J. Cui, F. Wen, and X. Tang, "Real Time Google and Live Image Search Re-Ranking," Proc. 16th ACM Int'l Conf. Multimedia, pp. 729-732, 2008.

[7] 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.

[8] H. Daume III and D. Marcu, "Domain Adaptation for Statistical Classifiers," J. Artificial Intelligence Research, vol. 26, pp. 101-126, 2006.

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

[10] B. Geng, L. Yang, C. Xu, and X.-S. Hua, "Ranking Model Adaptation for Domain-Specific Search," Proc. 18th ACM Conf. Information and Knowledge Management (CIKM '09), pp. 197-206, 2009.

[11] F. Girosi, M. Jones, and T. Poggio, "Regularization Theory and Neural Networks Architectures," Neural Computation, vol. 7, pp. 219-269, 1995.

[12] R. Herbrich, T. Graepel, and K. Obermayer, "Large Margin Rank Boundaries for Ordinal Regression," Advances in Large Margin Classifiers, pp. 115-132, MIT Press, 2000.

[13] K. Ja¨rvelin and J. Keka¨la¨inen, "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, 2000.

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

[15] T. Joachims, "Training Linear Svms in Linear Time," Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '06), pp. 217-226, 2006.

[16] M.G. Kendall, "A New Measure of Rank Correlation," Biometrika, vol. 30, nos. 1/2, pp. 81-93, June 1938.

[17] J.M. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "The Web as a Graph: Measurements, Models and Methods," Proc. Int'l Conf. Combinatorics and Computing, pp. 1-18, 1999.

[18] 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

[19] J. Lafferty and C. Zhai, "Document Language Models, Query Models, and Risk Minimization for Information Retrieval," Proc. 24th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '01), pp. 111-119, 2001.

[20] T.-Y. Liu, T. Qin, J. Xu, W. Xiong, and H. Li, "Benchmark Dataset for Research on Learning to Rank for Information Retrieval," Proc. SIGIR Workshop Learning to Rank for Information Retrieval (LR4IR '07), 2007.

[21] L. Page, S. Brin, R. Motwani, and T. Winograd, "The Pagerank Citation Ranking: Bringing Order to the Web," technical report, Stanford Univ., 1998.

[22] J.C. Platt, "Fast Training of Support Vector Machines Using Sequential Minimal Optimization," Advances in Kernel Methods: Support Vector Learning, pp. 185-208, MIT Press, 1999.

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

[24] S. Robertson and D.A. Hull, "The Trec-9 Filtering Track Final Report," Proc. Ninth Text Retrieval Conf., pp. 25-40, 2000.

[25] H. Shimodaira, "Improving Predictive Inference Under Covariate Shift by Weighting the Log-Likelihood Function," J. Statistical Planning and Inference, vol. 90, no. 18, pp. 227-244, 2000.

[26] Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large Margin Methods for Structured and Interdependent Output Variables," J. Machine Learning Research, vol. 6, pp. 1453-1484, 2005.

[27] V.N. Vapnik, Statistical Learning Theory. Wiley-Interscience, 1998.

[28] J. Xu and H. Li, "Adarank: A Boosting Algorithm for Information Retrieval," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 391-398, 2007.

[29] J. Xu, T.Y. Liu, M. Lu, H. Li, and W.Y. Ma, "Directly Optimizing Evaluation Measures in Learning to Rank," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 107-114, 2008.

[30] J. Yang, R. Yan, and A.G. Hauptmann, "Cross-Domain Video Concept Detection Using Adaptive Svms," Proc. 15th Int'l Conf. Multimedia, pp. 188-197, 2007.

[31] Y. Yue, T. Finley, F. Radlinski, and T. Joachims, "A Support Vector Method for Optimizing Average Precision," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 271-278, 2007.

[32] B. Zadrozny, "Learning and Evaluating Classifiers Under Sample Selection Bias," Proc. 21st Int'l Conf. Machine Learning (ICML '04), p. 114, 2004.

[33] Shilpy Sharma-" Information Retrieval in Domain Specific Search Engine with Machine Learning Approaches"- World Academy of Science, Engineering and Technology 18 2008.

[34] Andrew McCallum, Kamal Nigam, Jason Rennie and Kristie Seymore-" A Machine Learning Approach to Building Domain-Speci_c Search Engines".

[35] Sukanta Sinha, Rana Dattagupta, and Debajyoti Mukhopadhyay-" Web-Page Prediction for Domain Specific Web-Search Using Boolean Bit Mask".

[36] Satoshi Oyama, Takashi Kokubo, and Toru Ishida-"Domain-Specific Web Search with Keyword Spices"- IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 16, NO. 1, JANUARY 2004.

[37] Chunxia YIN, Jian LIU, Chao YANG, and Huiying ZHANG-" A Novel Method for Crawler in Domain-specific Search"- Journal of Computational Information Systems5:6(2009) 1749-1755- Available at http://www.JofCI.org

[38] Sofia Ceppi-" Designing Sponsored Search Auctions for Federated Domain-Specific Search Engines"