Scientific Journal Impact Factor (SJIF) : 1.711
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745
ISSN (PRINT) : 2393 - 8161

Expert Finding using discriminative infinite Hidden Markov Model

Mr.Yogesh.M.Here[1], Mr.H.A.Tirmare[2]
[1] *Department of Technology, Shivaji University Kolhapur*
[2]Assistant *Professor ,Department of Technology, Shivaji University Kolhapur*

Abstract—Process of finding the right expert for a given problem in an organization is becoming feasible. Using web surfing data it is feasible to find advisor who is most likely possessing the desired piece of fine grained knowledge related with given query. Web surfing data is clustered into tasks by using Gaussian Dirichlet process mixture model. In order to mine micro aspects in each task a novel discriminative infinite Hidden Markov Model is developed. The fine grained knowledge for each task can have hierarchical structure. In order to implement hierarchy apply the discriminative infinite Hidden Markov Model on micro aspects iteratively.

Keywords—Advisor Search, Gaussian Dirichlet process mixture model, discriminative infinite Hidden Markov Model, Micro aspects , task, web surfing data, Clustering

I. INTRODUCTION

In a collaborative environment every person shares information with every other person. It is common that different members of the group may try to access the same information separately. For example, this happens in a research lab, where members are focused on projects which require similar background knowledge. It may happen that the researcher tries to solve a particular problem using one method which he/ she is not familiar with but has been studied by another researcher. In this case finding the right individual who already knows something in that field is always superior than studying by oneself. Because people can provide live interaction, better communication and give more insights on that particular problem and also gives what were the problems faced by him/her. Finding the right person is always hard due to a number of reasons.

This scenario is different from conventional expert [14] search problem in that finding expert [16] on particular problem is dependent on associated documents in the enterprise repository. Our goal is to search advisors who know something related to that query. In order to analyze knowledge acquired by web users we will analyze users web surfing [1] and browsing activities which will reveal users knowledge gaining process on micro aspect level. For finding web surfing activities of each user one can use tcpdump for Linux platform and windump for Windows platform.Latent semantic structures in the Web surfing shows people's knowledge gaining process and web surfing data is good improvement over documents.

A two step framework for mining micro aspect in each task has been proposed. In the first step, Infinite Gaussian mixture model [3] based on Dirichlet Process [4] has been designed to cluster sessions which are generated by Web surfing activities. In the second step, micro aspect from session in each task has been extracted. A novel **d**iscriminative **i**nfinite **H**idden **M**arkov **M**odel (d-iHMM) is applied to the micro aspects for mining and patterns in each task. Finally a language model [19] based expert search method is applied over mined aspects to search for advisor. The aim of topic is not to find persons who are experts but to find persons who have desired knowledge related to the query. Also implementing fine-grained knowledge in hierarchical way is difficult because knowledge can contain micro aspects with similar topics. This problem can be solved by applying d-iHMM model iteratively. In order to solve d-iHMM Beam sampling [5] is used.

II. LITERATURE SURVEY

Ziyu Guan, Shengqi Yang, Huan Sun, Mudhakar Srivatsa, and Xifeng Yan [1] have proposed "Fine-Grained Knowledge Sharing in Collaborative Environments ". In paper they explains new method of expert finding and implement it. They find out expert using web surfing and browsing contents. Web surfing data gives more accurate results than traditional document based method. Gaussian Dirichlet process mixture model is used for clustering session in each task. In order to implement to mine micro aspects in each task discriminative infinite Hidden Markov Model is used.

Krisztian Balog [2] and Group have presented paper on "Formal models for expert finding in enterprise Corpora". Two general strategies have been presented to expert searching given a collection of document. The first directly models an expert's knowledge based on the documents that they are associated with and while the second locates documents on the queried topic and then finds the associated expert.

Carl Edward Rasussen [3] explained "The Infinite Gaussian Mixture Model ". In which author show that how infinite mixture model has several advantages over finite mixture model. Infinite mixture model achieves good performance on multi dimensional data. It is simple to handle infinite limit than than to work with finite models with known sizes.

Jurgen Van Gael , Yunus Saatci ,Yee Whye The , Zoubin Ghahramani [5] given paper on "Beam Sampling for the Infinite Hidden Markov Model ". This paper presents how Beam sampling algorithm outperforms the Gibbs algorithm and more robust than Gibbs algorithm. They also show that Beam Sampler is faster than Gibbs algorithm and easy to implement.

Dawit Yimam [6] presents paper on "Expert finding System for Organizations: Domain analysis and the DEMOIR Approach". Author proposes Dynamic Expertise Modeling from Organizational Information Resources Server [7] approach to formulate expert finding system that meet the said requirements.

Finding experts [12] in collaborative environment to share information has lots of advantages, early approach to advisor search is that data about skills and knowledge of each individual in the organization is collected. This data is stored in the database manually, such an approach requires considerable effort to set up and maintain the data [10]. More recent techniques [9] judge the expert automatically. The task of expert search received significant importance since it has been implemented in the TREC enterprise track [11].

Macdonald and Ackerman [13] distinguish several aspects of expert finding. They call it as expert identification or advisor recognition ("Who are the experts on topic X ?") and expertise selection ("What does expert Y know ?").In expert identification query is passed related to the topic X and the relevant documents are generated as output. In the output the relevance of each document is found and those documents who have the highest relevance will be ranked high. From the highest rank document it will be easy to recognize advisor.

Yi Fang, Luo Si and Aditya P. Mathur [14] had published article on "Discriminative models of integrating document evidence and document-candidate associations for expert search". They proposed principle relevance based discriminative learning framework for expert search and derive specific discriminative models from the framework. The proposed research can naturally integrate various documental evidence and candidate-document associations into a single model without extra modeling assumptions or effort.

Pavel Serdyukov and team members [15] explained "Modeling multi-step relevance propagation for expert finding". This paper proposes a novel approach to expert finding in large enterprises or Intranets by modeling candidate experts, organizational documents and various relations among them with the so-called *expertise graphs.* They model the process of expert finding by probabilistic random walks of three kinds: finite, infinite and absorbing.

Witold Abramowicz, Elzbieta Bukowska, Monika Kaczmarek and Monika Starzecka [17] proposed paper on "Semantic enabled Efficient and Scalable Retrieval of Experts". There system proposes user friendly interface to perform queries that allow to find person with specific characteristics.

Krisztian Balog, Leif Azzopardi ,Maarten de Rijke [19] published a article on "A language modeling framework for expert finding". In which they introduce and detail language modeling approaches that integrate representation ,association and search of experts using various textual data sources into a generative probabilistic framework. They introduces two model in which first model finds out prominent topic in the document and in the second model they identify important documents for a given topic and determine who is most closely associated with these documents.

Hui Fang and ChengXiang Zhai [20] have proposed "Probabilistic Models for Expert Finding". In this paper they find persons who are knowledgeable about a given topic by general probabilistic framework in which they proposes two types of generative model that is candidate generation model and topic generation model.

Xiayong Liu, W.Bruce Croft ,Matthew Koll [18] have proposed "Finding Experts in Community Based Question –Answering Services". They have explained and analyzed different language models for finding experts. More specifically models used in their work is query likelihood model ,the relevance model and cluster based model. All these models give ranking to the peoples profile and higher ranked peoples are considered as experts.

III. CHOICE OF TOPIC WITH REASONING

If we simply treat web surfing data as collection of document and apply the traditional search method then candidate who have searched more data but which is not relevant to query will be ranked as high as compared to the candidate who have searched less but more relevant data. Hence this method of finding advisory gives wrong result.

In order to solve that problem we will analyze first ,candidates fine grained knowledge occurred in the web surfing data by using semantic structures and then search over fined pieces of fine grained knowledge, which will give correct advisors with more relevant contents. Semantic structures in the web will give peoples knowledge acquisition process in deep level. Our goal is to find out advisors who are most likely possessing desired fine grained knowledge.

Traditional hierarchical clustering methods using infinite Hidden Markov Model will not able to distinguish micro aspects of a task. In order solve that problem we are implementing discriminative infinite Hidden Markov Model which will not allow messing up of micro aspects in a task. For example "Java IO" can contain "File IO" and "Network IO" as subtasks. In order to implement the fine grained knowledge in hierarchical way apply the discriminative infinite Hidden Markov Model iteratively.

IV. OUTLINE OF WORK

The proposed system have following modules:

- **End user module :-**
  In Web surfing module, we propose to log and analyze users web surfing data (not only search but also browsing activities), which reveals a user's knowledge gaining process. Web surfing data provides more comprehensive information about the knowledge gaining activities of users. End user gives query to the search engine.

- **Session clustering module: -** The need for clustering is that contents generated for the same task may be textually similar while for the different task may be different. Hence, clustering is required for recovering tasks from sessions. Clustering of sessions is done by Infinite Gaussian mixture model based on Dirichlet process.

- **Micro aspect module: -** Micro aspects in the task are already similar with another. If we model each component independently it is likely that we mess up sessions from different micro aspect. Therefore we should model different micro aspects in a task jointly, separating the common content characteristics of the task from the distinctive characteristics of each micro aspect. Novel discriminative infinite hidden Markov Model (d-iHMM) [3] is used to mine micro aspects and possible evolution patterns in a task. Beam sampling [5] is used for solving d-iHMM model.

- **Advisory search module:-** Advisor search is dedicated to retrieving people who have desired piece of fine grained knowledge. Here for finding advisor search we will use web sessions instead of documents. Because sessions will give accessed information in micro aspect level. Three advisor search schemes are compared such as session based, micro aspect based and task based with increasing granularity. Micro aspects based scheme outperforms the session based scheme and task based scheme. In order to get advisory we apply a language model based expert search method.

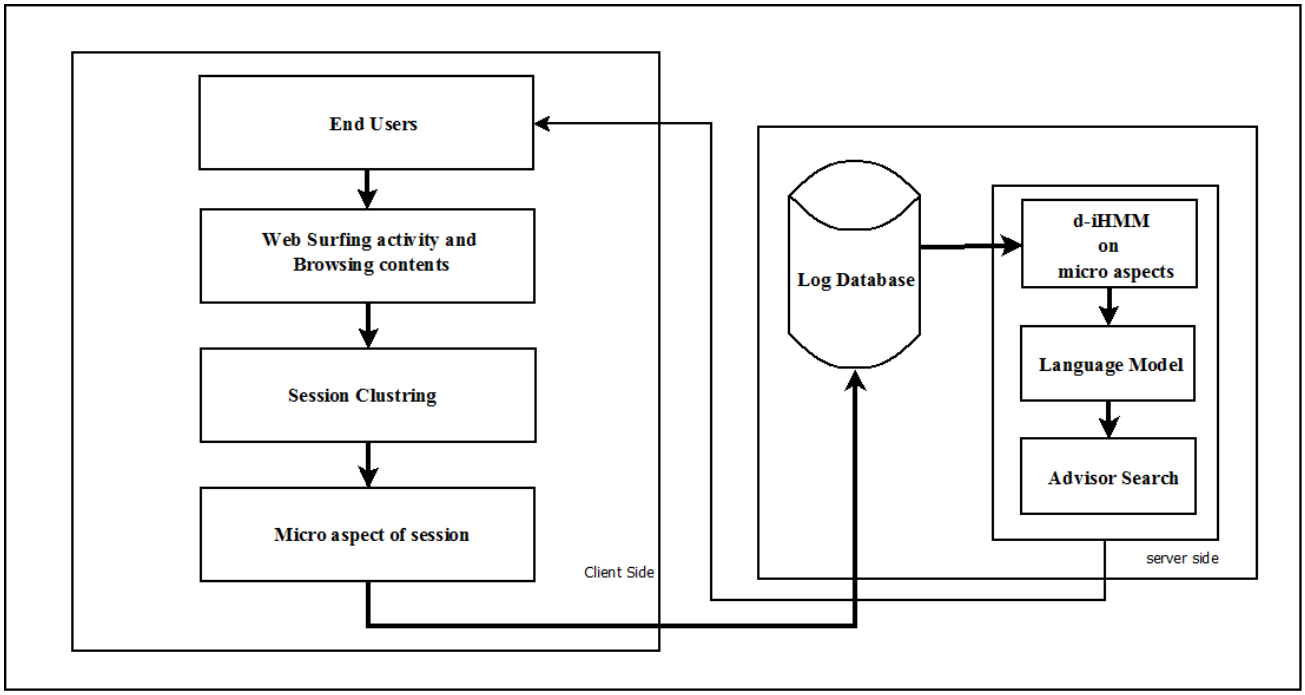

Figure : System Architecture

V. EXPERT SEARCH

Expert Search is concerned with finding experts on a particular topic. This problem has many real world application. Currently, people have to manually discover the experts, which is obviously

labor intensive and time overwhelming. Thus, it would be very interesting to study how to automatically identify experts for a specified expertise area. Early approaches involve building a knowledge base which contains the descriptions of people's skills within an organization. Expert search became a hot research area since the start of the TREC enterprise track in 2005.The proposed advisor search problem is different from traditional expert search. (1) Advisor search is dedicated to retrieving people who are most likely possessing the desired piece of fine-grained knowledge, while traditional expert search does not explicitly take this goal. (2) The critical difference lies in the data, i.e. sessions are significantly diverse from documents in enterprise repositories. A person typically generates multiple sessions for a micro aspect of a task, e.g. a person could spend many sessions learning about Java multithreading skills. In other words, the uniqueness of sessions is that they contain semantic structures which reflect people's knowledge acquisition process. If we treat sessions as documents in an enterprise repository and apply the traditional expert search methods, we could get incorrect ranking: due to the accumulation nature of traditional methods, a candidate who generated a lot of marginally relevant sessions will be ranked higher than the one who generated less but highly relevant sessions, for the query "Java multi-thread programming" . Therefore, it is important to recognize the semantic structures and summarize the session data into micro-aspects so that we can find the desired advisor accurately. In this paper we develop nonparametric generative models to mine micro aspects and show the superiority of our search scheme over the simple idea of applying traditional expert search methods on session data directly. This line of research tries to recover tasks from people's search behaviors and bears some similarity to our work. Nevertheless, our work differs from theirs from the following aspects. First, we consider general web surfing contents, rather than search engine query logs. Query logs do not record the subsequent surfing activity after the user clicked a relevant search result. Moreover, it is found that 50 percent of a user's online page views are content browsing [17]. Web surfing data provides more comprehensive information about the knowledge gaining activities of users. Although various methods were proposed for extracting search tasks in query logs, these methods cannot be applied in our setting since they exploit query log specific properties. Second, none of the above works tried to mine fine-grained aspects for each task. When studying, people could spend some effort on one fine-grained aspect of a task and generate multiple contents. Summarizing fine-grained aspects can provide a fine grained description of the knowledge gained by a person. Finally, none of existing works which analyze user online behaviors tried to address advisor search by exploiting the data generated from users' past online behaviors.

VI. MINING MICROASPECTS

The major challenge of mining micro aspects is that the micro aspects in a task are already similar with one another. If we model each component (i.e. micro-aspect) independently, it is likely that we mess up sessions from different micro-aspects, i.e. leading to bad discrimination. Therefore, we should model different micro-aspects in a task jointly, separating the common content characteristics of the task from the distinctive characteristics of each micro-aspect. To this end, we extends the infinite Hidden Markov Model (iHMM) and propose a novel discriminative infinite Hidden Markov Model to mine micro-aspects and possible evolution patterns in a task.

After we obtain the mined micro aspects of each task, advisor search can then be implemented on the collection of learned micro aspects. We apply the traditional language model based expert search method.

We first show the results of advisor search. Three schemes are compared: session-based, micro-aspect-based and task based with an increasing granularity. The language model based expert search method mentioned is used as the retrieval method. We have tried using other traditional expert search methods, but the results are very similar since they all intrinsically accumulate relevance

scores of associated "documents" to candidates. For each scheme, a language model is constructed for each "document", i.e. a session, a micro-aspect, or a task, by aggregating all the texts belonging to it. Note that the session-based scheme is intrinsically applying the traditional language model based expert search method on web surfing data directly.

VII. CONCLUSION

We introduced new technique for finding expert in collaborative environment. Finding the right person in an organization with the appropriate skills and knowledge is often crucial to the success of projects being undertaken. In order to find right person we proposed novel discriminative infinite hidden Markov Model to mine micro aspects and evolution pattern of each task. This method of finding expert is better than traditional expert search problem where expert search aims to find domain experts based on associated documents.

REFERENCES

[1] Ziyu Guan, Shengqi Yang, Huan Sun, Mudhakar Srivatsa, and Xifeng Yan "Fine-Grained Knowledge Sharing in Collaborative Environments". IEEE transactions on Knowledge and data Engineering, vol.27,No. 8,August 2015, pp.2163-2174.

[2] K. Balog, L. Azzopardi, and M. de Rijke, "Formal models for expert finding in enterprise corpora," in Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2006, pp. 43–50.

[3] J. Van Gael, Y. Saatci, Y. Teh, and Z. Ghahramani, "Beam sampling for the infinite hidden Markov model," in Proc. Int. Conf. Mach. Learn., 2008, pp. 1088–1095.

[4] Rasmussen, "The infinite Gaussian mixture model," in Proc. Adv. Neural Inf. Process. Syst., 2000, pp. 554–560.

[5] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical Dirichlet processes," J. Am. Statist. Assoc., vol. 101, no. 476, pp. 1566–1581, 2006

[6] D. Yimam. Expert finding systems for organizations: Domain analysis and the demoir approach. In ECSCW 999 Workshop: Beyond Knowledge Management: Managing Expertise, pages 276–283, New York, NY, USA, 1996. ACM Press

[7] D. Yimam-Seid and A. Kobsa. Expert finding systems for organizations: Problem and domain analysis and the demoir approach. Journal of Organizational Computing and Electronic Commerce, 13(1):1–24, 2003.

[8] A. Mockus and J. D. Herbsleb. Expertise browser: a quantitative approach to identifying expertise. In ICSE '02: Proceedings of the 24th International Conference on Software Engineering, pages 503–512. ACM Press, 2002.

[9] T. H. Davenport and L. Prusak. Working Knowledge: How Organizations Manage What They Know. Harvard Business School Press, Boston, MA, 1998.

[10] N. Craswell, A. de Vries, and I. Soboroff. Overview of the trec-2005 enterprise track. In *TREC-13*, 2005

[11] D.Yimam-Seid and A. Kobsa. Expert finding systems for organizations. Sharing Expertise: Beyond Knowledge Management, 2003

[12] D. W.McDonald and M. S. Ackerman. Expertise recommender: a flexible recommendation system and architecture.In CSCW '00: Proceedings of the 2000 ACM conference on Computer supported cooperative work, pages 231–240. ACM Press, 2000.

[13] Y. Fang, L. Si, and A. P. Mathur, "Discriminative models of integrating document evidence and document-candidate associations for expert search," in Proc. 33rd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2010, pp. 683–690.

[14] P. Serdyukov, H. Rode, and D. Hiemstra, "Modeling multi-step relevance propagation for expert finding," in Proc. 17th ACM Conf. Inf. Knowl. Manage., 2008, pp. 1133–1142.

[15] https://en.wikipedia.org./wiki/Expertise_finding

[16] Witold Abramowicz, Elzbieta Bukowska,Monika Kaczmarek and Monika Starzecka "Semantic enabled Efficient and Scalable Retrieval of Experts".

[17] X. Liu, W. B. Croft, and M. Koll, "Finding experts in community based question-answering services," in Proc. 14th ACM Int. Conf.Inf. Knowl. Manage., 2005, pp. 315–316.

[18] Krisztian Balog, Leif Azzopardi ,Maarten de Rijke "A language modeling framework for expert finding".www.elsevier.com/locate/infoproman Information processing and management,June 2008.

Hui Fang and ChengXiang Zhai "Probabilistic Models for Expert Finding".29th European Conference on IR Research ,ECIR 2007,Rome,Italy,April 2-5,2007,pp.418-430.