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
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 10, October 2017, pg.54 – 58

Providing Advisor Search with Fine Grained Knowledge Sharing

Tejas Sanghrajka[1], Dr. V. C. Kotak[2]

[1]Department of Information Technology, Shah & Anchor Kutchhi Engineering College, Mumbai, India

[2]Department of Electronics, Shah & Anchor Kutchhi Engineering College, Mumbai, India

[1] tejas.sanghrajka07@gmail.com

Abstract— To overcome all problems of not getting relevant information knowledge sharing is the remedy. Knowledge sharing can be done by analyzing knowledge acquired by users. In order to analyze knowledge acquired by web users, analysis of user's web surfing data is very useful. Using web surfing data it is possible to find advisor who most likely possesses the desired piece of fine grained knowledge related with given query. In this dissertation work investigation is done on fine-grained knowledge sharing in collaborative environments. In this work a methodology is proposed to analyze member's web surfing data to summarize the fine-grained knowledge acquired by them and keep a record. A two-step framework is proposed for mining fine-grained knowledge: (1) web surfing data is clustered into tasks by a nonparametric generative model; (2) a novel discriminative infinite Hidden Markov Model (d-iHMM) is developed to mine fine-grained aspects in each task. Finally, the classic expert search method is applied to the mined results to find proper advisor for knowledge sharing. The proposed system is described, and its performance is also evaluated.

Keywords— Web surfing data, Clustering, knowledge sharing, discriminative infinite Hidden Markov Model, fine-grained knowledge, nonparametric generative model.

I. INTRODUCTION

Most powerful feature of the web, however, is the wealth of information it contains, much of which is not available in even the largest library collections. Information about almost any subject is available in depth and up to date. This is incredibly valuable for every subject you can imagine. It facilitates rapid interactive communication and the exchange of huge volumes of data.

But similarly it also raises the complexity of how to deal with the information from the different perspectives of view – users, Web service providers, business analysts. With the exponential growth of WWW, it has become difficult to access desired information that matches with user needs and interest. It requires an efficient information search strategy otherwise it causes danger of overload and excess information or the search can be slowed down. There is a lot of apparently unconnected information so it is difficult to filter and prioritize information. There is no guarantee of finding what one is looking for.

To overcome these all problems of not getting relevant information knowledge sharing is the remedy. Knowledge sharing helps address this problem. Just as the recycling of materials is good for the environment, reuse of information is good because it minimizes rework, prevents problems, saves time, and accelerates progress.

In a collaborative environment, it could be common that members try to acquire similar information on the web in order to gain specific knowledge in one domain. For example, Alice wants to download a dictionary and she starts to surf the web and she doesn't get the direct download link which has already been accessed by Bob. In this case, it might be a good idea to consult Bob or go through the links accessed by Bob, rather than searching by herself and wasting time on useless links redirecting to the another page.

This dissertation presents a novel method to identify, how to enable such knowledge sharing mechanism by analyzing user data. The goal of this proposed work is to find proper "advisors" who are most likely possessing the desired piece of fine-grained knowledge based on their web surfing activities. This work departs from the traditional expert search method in that expert search aims to find domain experts based on their associated documents in an enterprise repository. In order to analyze the knowledge acquired by web users, new method is proposed to log and analyze user's web surfing data. User's interactions with the web can be segmented into different "tasks", e.g., "web mining" and "shopping". Textual contents of a task are usually cohesive. A task can be further decomposed into fine-grained aspects (called micro-aspects). A micro-aspect could be roughly defined as a significantly more cohesive subset of sessions in a task. For example, the task "web mining" might contain "usage mining" and "content mining". To this end, a novel discriminative infinite Hidden Markov Model (d-iHMM) is proposed to mine micro-aspects in each task. Finally, a language model based expert search method is applied over the mined micro aspects for advisor search. Also if any person wants to keep his name secret about any topic then the privacy can also be provided. Without showing his name the search made by him can be suggested to another person. This can give a proper direction to ones search.

II. LITERATURE REVIEW & RELATED WORK

In year 2010 David M.Ble, Thomas L.Griffiths and Michael I.Jordan [6] presented the nested Chinese restaurant process (nCRP), that assigns probability distributions to infinitely deep branching trees. Proposed Bayesian nonparametric model is based on representations that are allowed to grow structurally as more data are observed.

But analyzing the richly structured data requires extending this approach.

In year 2008 Matthew D. Hoffmany, Perry R. Cookyz, David M. Bleiy [5]explained the markove process and the hidden markove model. They considered that each successive state still depends only on the state(s) before it, but we cannot observe these states directly these are hidden states.

But it restricts the number of hidden states to grow as data grows infinitely. There is a need of infinite approach.

In year 2001 M. J. Beal, Z. Ghahramani, and C. E. Rasmussen [1] had shown that it is possible to extend hidden Markov models to have a countably infinite number of hidden states. By using the theory of Dirichlet processes they implicitly integrated out the infinitely many transition parameters.

But the major challenge of mining micro aspects is that the micro aspects in a task are already similar with one another. It mess up sessions from different micro-aspects, i.e. leading to bad discrimination

In year 2003 D. M. Blei, A. Y. Ng, and M. I. Jordan [2] had analyzed topic modeling. Topic modeling is a popular tool for analyzing topics in a document collection. The most prevalent topic modeling method is Latent Dirichlet Allocation (LDA). It is a generative probabilistic model for collections of discrete data. Topic modeling decomposes a document into topics.

But it doesn't recover the semantic structures of people's online learning activities from their web surfing data, i.e. identifying groups of sessions representing tasks (e.g. learning "Java") and micro-aspects (e.g. learning "Java multithreading"). After applying topic modeling methods on session data, it is still difficult to find the right advisor by using the mined topics.

In year 2005 X. Liu, W. B. Croft, and M. Koll [3] has also been studied expert retrieval in other scenarios, e.g. online question answering communities. People using such services are like a community – anyone can ask, anyone can answer, and everyone can share, since all of the questions and answers are public and searchable immediately.

But there are hundreds of questions asked each day but some portion of them may not be answered or there may be a lag between the time when a question is asked and when it is answered. Also the answers may not be satisfactory.

In year 2006 K. Balog, L. Azzopardi, and M. de Rijke [4] proposed a language model framework for expert search. Expert search aims at retrieving people who have expertise on the given query topic. Their Model 2 is a document-centric approach which first computes the relevance of documents to a query and then accumulates

for each candidate the relevance scores of the documents that are associated with the candidate. It locates documents on topic, and then finds the associated expert. Balog showed that Model 2 performed better.

But the nature of these methods is still accumulating relevance scores of associated documents to candidates. Traditional expert search does not explicitly retrieving people who are most likely possessing the desired piece of fine-grained knowledge it focused on finding experts only rather than to mine fine-grained aspects for each task.

In year 2010 A. K. Jain[7] had given the idea of k-means algorithm. k-means algorithm is explained for clustering the data.

But it allows only hard assignments which restricts the data and doesn't allow to express uncertainty.

In year 2011 A. Kotov, P. Bennett, R. White, S. Dumais, and J. Teevan [8] designed classifiers to identify same-task queries for a given query and to predict whether a user will resume a task. They introduced and addressed the two problems in the context of analysis of cross-session search tasks: (i) identifying queries from earlier sessions on the same task, and (ii) predicting whether a user will return to the same task during a later session.

But it doesn't provided richer prediction models and alternative feature sets, exploring new prediction and classification problems in the context of cross session information needs. It also didn't tried to mine fine-grained aspects for each task. Summarizing fine-grained aspects can provide a fine-grained description of the knowledge gained by a person.

In year 2013 Zoubin Ghahramani[9] defined An infinite Gaussian mixture model for clustering. It uses hard assignment and allows to express uncertainty.

But there is still a non parametric approach is needed for finding number of clusters.

In year 2015 Ziyu Guan, Shengqi Yang, Huan Sun, Mudhakar Srivatsa, and Xifeng Yan [10] suggested a fine-grained knowledge sharing in collaborative environments. They proposed a method to find proper "advisors" who are most likely possessing the desired piece of fine-grained knowledge based on their web surfing activities.

But the fine-grained knowledge could have a hierarchical structure. And how to search over this hierarchy is not a trivial problem. Also this work creates an issue of privacy.

III. PROBLEM DEFINITION

The creation of knowledge is needed to be aligned with sharing activities. There is a need of sharing knowledge and working together to accomplish stated goals and objectives. Approaching to a right person can be far more efficient than studying by oneself, since people can provide digested information, insights and live interactions, compared to the web.

The problems are:

- To analyze user online behaviors to mine the tasks and provide advisor search.
- To provide ease of access of desired information and save time of repetitive efforts.
- To provide the results to the user according to usage made by user.
- Many times people don't get direct links or proper direction to absorb the actual data they want.
- Many people pointed out the problem of finding a right advisor due to the variety of information needs.
- Repeating efforts occurs when people try to acquire similar information on the web in order to gain specific knowledge in one domain.
- These repeating efforts cause waste of time and energy.
- Users are not getting information sharing environment which can enhance their knowledge.
- Studying on the web is not as efficient as learning from an advisor and from the accessed results.

IV. OBJECTIVES

- To retrieve people who are most likely possessing the desired piece of fine-grained knowledge.
- To summarize fine-grained aspects which can provide a fine grained description of the knowledge gained by a person.
- To provide a healthy sharing environment which can enhance their knowledge
- To retrieve previously accessed data by another users to perform knowledge sharing.

V. PROPOSED SYSTEM

The goal of this proposed work is to find proper "advisors" who are most likely possessing the desired piece of fine-grained knowledge based on their web surfing activities. This work provides the advisor and the accessed data of that advisor to the user to make the search operation more productive and easy. The goal of this method is not finding domain experts but a person who has the desired piece of knowledge. This methodology provides a way to find proper "advisor" who are most likely possessing the desired piece of fine-grained knowledge based on their web surfing activities. This work proposes the fine-grained knowledge sharing in collaborative environments. This method is proposed to solve the problems by first summarizing web surfing data into fine grained aspects, and then search over these aspects. To this end, a novel discriminative infinite Hidden Markov Model (d-iHMM) is proposed to mine micro-aspects in each task. Also if any person wants to keep his name secret about any topic then the privacy can also be provided. Without showing his name the search made by him can be suggested to another person. This can give a proper direction to ones search. A real user-generated web surfing data is collected to test the feasibility of this method. Then the mined micro-aspects of each task is obtained, advisor search can then be implemented on the collection of learned micro-aspects. Each member is first registered and provided a unique ID which helps to keep record of member's search data and helps in advisor search.

VI. PROPOSED SYSTEM ARCHITECTURE

The architecture is explained step by step as follows:

Step 1: New user is registered first and then performs login.

Step 2: User is checked for validation. After checked for validation the user enters the query.

Step 3: Here user is asked for privacy whether to click it or not depends on user.

Step 4: If privacy is clicked yes then the database stores urls of that query only but if privacy is clicked no then it stores name along with the urls.

Step 5: Then query is passed to the search engine and similarly checked in database for finding match.

Step 6: User entered web surfing data including queries and user detail is analyzed and saved.

Step 7: if match found to the query entered then it proceeds further and if not then it shows no search done previously.

Step 8: The web surfing data is clustered into tasks by a nonparametric generative model.

Step 9: These tasks can be further decomposed into fine-grained aspects (called micro-aspects).

Step 10: Then discriminative infinite Hidden Markov Model is developed to mine fine-grained aspects in each task.

Step 11: Finally, a language model based expert search method is applied over the mined micro aspects for advisor search.

Step 12: The web page display with the advisor search is provided to the end user along with the previously accessed data and the urls clicked privacy as the reference to the user.

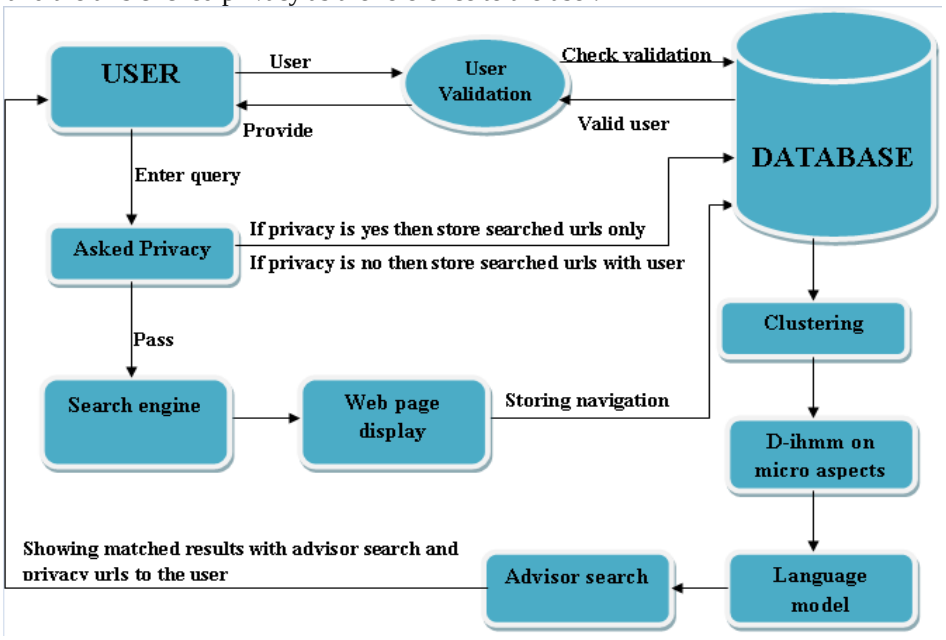

Figure 1: ARCHITECTURE DIAGRAM

VII. COMPARISON WITH EXISTING SYSTEMS

- The proposed work departs from the traditional expert search method in that expert search aims to find domain experts based on their associated documents in an enterprise repository whereas the proposed method provides advisor search based on analyzing user accessed web surfing data.
- Traditional method uses k means algorithm for clustering and hidden markove model is used for mining microaspects.
- Whereas the proposed method make use of Gaussian mixture model for clustering user accessed data and adopt the non parametric approach for number of clusters and uses discriminative infinite hidden markove model for mining microaspects and then provide advisor search by using language model.
- This work overcomes the problem of maintaining privacy of user which is not solved in the existing methods.

VIII. CONCLUSION

Among the various techniques proposed for clustering of search results, as described in the literature review, combination of Gaussian mixture model and non parametric generative model is selected and then the discriminative infinite hidden markove model is used for the implementation of the project to provide fine grained information to the user.

It provides a healthy sharing environment which can enhance people knowledge. It provides an easy way to retrieve people who are most likely possessing the desired piece of fine-grained knowledge by addressing advisor search by exploiting the data generated from user's past online behaviours. It provides access to the previously accessed search of users to make data easily available. It provides ease of access of desired information and save time of repetitive efforts. Also identified digging out fine-grained knowledge reflected by people's interactions with the outside world as the key to solving the problems. This method proposed a two-step framework to mine fine-grained knowledge and integrated it with the classic expert search method for finding right advisors.

IX. FUTURE WORK

The work in this dissertation can be implied best in a collaborative environment, where it is common that members try to acquire similar information on the web in order to gain specific knowledge in one domain and can be extended further for use on large scale. Also more semantic approach is needed to be discovered

ACKNOWLEDGEMENT

My thanks to the Guide, Dr. V. C. Kotak**,** who provided me constructive and positive feedback during the preparation of this paper.

REFERENCES

[1] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen, "The infinite hidden Markov model," in Proc. Adv. Neural Inf. Process. Syst., 2001, pp. 577–584.
[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.
[3] X. Liu, W. B. Croft, and M. Koll, "Finding experts in communitybased question-answering services," in Proc. 14th ACM Int. Conf. Inf. Knowl. Manage., 2005, pp. 315–316.
[4] 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.
[5]Matthew D. Hoffmany, Perry R. Cookyz, David M. Bleiy "DATA-DRIVEN RECOMPOSITION USING THE HIERARCHICAL DIRICHLET PROCESS HIDDEN MARKOV MODEL", 2008.
[6] David M.Blei, Thomas L.Griffiths,Michael I.Jordan," The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies" in Journal of the ACM, Vol.57, No.2, Article7, January 2010.
[7]A. K. Jain, "Data clustering: 50 years beyond k-means," Pattern Recog. Lett., vol. 31, no. 8, pp. 651–666, 2010.
[8] A. Kotov, P. Bennett, R. White, S. Dumais, and J. Teevan, "Modeling and analysis of cross-session search tasks," in Proc. 34th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2011, pp. 5–14.
[9]Zoubin Ghahramani, "Bayesian non-parametrics and the probabilistic approach to modeling", rsta.royalsocietypublishing.org,2013.
[10] Ziyu Guan; Shengqi Yang; Huan Sun; Srivatsa, M.; Xifeng Yan, "Fine-Grained Knowledge Sharing in Collaborative Environments," in Knowledge and Data Engineering, IEEE Transactions on , vol.27, no.8, pp.2163-2174, Aug. 1 2015

