Using Baum-Welch Algorithm for Sharing Fine-Grained Knowledge in Mutual Environments

A.M.Pushpalatha
PG Scholar
Department of Computer Science & Engineering
Christian College of Engg and Tech Dindigul,Tamilnadu-624619,India

Dr.A.Nirmal Kumar
Assistant Professor
Department of Computer Science & Engineering
Christian College of Engg and Tech Dindigul,Tamilnadu-624619,India

Abstract

Knowledge Sharing is an activity through which knowledge is exchanged among people, friends, families, communities or organizations. Mutual Environments, which enable company-wide global teams to identify the source of the antidote to a lack of preparedness. This paper investigates Fine grained knowledge sharing in collaborative environments.Two step framework is used. 1) Web surfing data are clustered into tasks by LEGDP (LaplacianEigenmap Gaussian DirichletProcess) Model. 2) From Each Task Micro Aspects are extracted by d-iHMM (Discriminative-infinite Hidden Markov Model) model. And to find proper members for knowledge sharing, the expert search method is applied on the mined results. Existing Hidden Markov Models takes larger memory and execution time.Also it provides low accuracy results. To overcome this we propose Baum-Welch algorithm.This algorithm is the extension of Hidden Markov Model. This provides more accuracy than HMM and also takes less execution time to find the best advisor for our related query.

Keywords: User Request, Traditional Expert Search, Preprocessing, LEGDP, d-iHMM, Advisor Search

I. INTRODUCTION

Data Mining is the thetaking out of secretedanalytical information from huge databases. Itis a powerful new technology and helps to companiesfor focus on the most important information in their data warehouses. It is the process of analyzing data from different perspectives and summarizing that into useful information. Data Mining is also called as data or knowledge discovery. In this project dataset is created based on the user request. This clusters all web surfing data into tasks.

Web Mining is the application of data mining to discover patterns from the World Wide Web(WWW).Web Mining has three types. Web Usage mining is used to discover interesting usage patterns from the web data. Usage data captures the identity of web users along with their browsing behavior at the web site. Web structure Mining is the process of analyzing node and connection structure of web site. Web Content Mining is the extraction and integration of useful information and knowledge from Web page content. Data Mining applications are ranging from commercial to social domains. Present-Day data mining is a progressive multidisciplinary work. This multidisciplinary approach is well reflected within the field of information systems.

II. RELATED WORKS

My work is closely related to several group of research works.

KrisztianBalog, Leif Azzopardi and Maarten de Rijke (2006)have proposed the Formal Models for Expert Finding in Enterprise Corpora [3].Searching an organization's document repositories for experts andprovides a cost effective resolution to the task of expert finding. They present two general strategies for expert searchingand give a collection of documents. These are dignified using multiplicative probabilistic models. The first of these directly models are based on the documents and they are associated with the second topic of documents and then finds the associated expert. Forming reliable associations are crucial to the performance of expert finding systems. Consequently, in this evaluation they compare the different approaches and then exploring a variety of associations along with other operational parameters (such as topicality). Using the TREC Enterprise corpora, they show that the second strategy consistently outperforms the first. A comparison against other unsupervised techniques, reveals that their second model delivers excellent performance.

Latent Dirichlet Allocation [4]was proposed by David Blei M and Michael Jordan I (2003).Topic modeling is a popular tool for analyzing topics in a document collection. The most prevalent topic modeling method is Latent Dirichlet Allocation (LDA) . Based on LDA, various topic modeling methods have been proposed, e.g. the dynamic topic model for sequential data and the hierarchical topic model for building topic hierarchies. The Hierarchical DP (HDP) model can also be instantiated as a nonparametric version of LDA .

Infinite Hidden Markov Model [5] work is proposed by Carl Edward Ramuseen and Mathew Beal J (2004).The discriminative-infinite Hidden Markov Model (d-iHMM) is used for mining micro-aspects from each task. The major task of mining micro-aspects is that the micro-aspects from tasks are already similar with one another. If they model each component (i.e. micro-aspect) independently (as most traditional models do), it is likely that they mess up sessions from different micro-

aspects, i.e. leading to bad discrimination. Therefore, they should model different micro-aspects in a task jointly and separating the common content characteristics of the task from the distinctive characteristics of each micro-aspect. To this end, they extends the infinite Hidden Markov Model (iHMM) andfor mining micro-aspects, propose a novel discriminative infinite Hidden Markov Model and also for possible evolution patterns in each task.

To extend hidden Markov models,this is possible to have a accountably infinite number of hidden states. They can indirectlyassimilate the infinitely many transition parameters anddeparture only three hyperparameters by using the theory of Dirichlet processes which can be learned from data. Three of these hyperparameters define a hierarchical Dirichlet process and they arecapable of capturing a rich set of transition dynamics.

David BleiM and Michael Jordan I (2006)was proposed The Variational Inference for Dirichlet Process Mixtures [6] .Dirichlet process (DP) mixture models are the cornerstone of non-parametric Bayesian statistics and the development of Monte-Carlo Markov chain(MCMC) sampling methods for DP mixtures has enabled the application of non- parametric Bayesian methods to a variety of practical data analysis problems. Though, MCMC sampling can be excessively slow and it is important to explore alternatives. By variational methods, one class of alternatives are provided a class of deterministic algorithms that convert inference problems into optimization problems.

Discriminative Models of Mixing Document Evidence and Document-Candidate Associations [7] for Expert Searchpaper was proposed by Aditya P. Mathur, Luo Si and Yi Fang (2010).Generative models such as statistical language modeling hasbeen widely studied in the task of expert search to model the relationship between experts and their knowledge, which is indicated in supporting documents. On the other hand, discriminative models are received little attention in expert search examination, although they have been shown to outperform generative models in many other information retrieval and machine learning applications. For expert search, this paper propose a principled relevance-based discriminative learning framework and it derives specific discriminative models from that framework.

Data clustering: 50 years beyond K-means [8] was proposed by Anil K. Jain (2010).Consolidating data into sensible groupings is called Clustering.The objective of clustering is to find the structure ofdata.One of the most popular and simple clustering algorithm is K-means algorithm. Cluster analysis is the formal study of methods andalgorithms for grouping, or clustering, objects allowing to measured or perceived intrinsic characteristicsor similarity. Cluster analysis does not use category labels, that tabitems with former identifiers, i.e.,class labels. The non-appearance of category information is used to differentiate data clustering (unsupervised learning) from classification or discriminant analysis (supervised learning).

Hierarchical Clustering Algorithms for Document Datasets [9] work was proposed by yingzhao and George karypis (2005).By organizing large amounts of information into a small number of meaningful clusters,fast and high-quality document grouping algorithms are played asignificant role for providing intuitivenavigation and browsing appliances. In particular, for large document collections, clustering algorithms arebuilding meaningful hierarchies and that are ideal tool for their collaborativeconception and investigation.They provide data-views, that are consistent, predictable, and different levels of granularity. These paper emphases on document clustering algorithms that buildsuch hierarchical resolutions.

A Characterization of Online Search Behavior [11] paper was proposed by Ravi Kumar Andrew Tomkins (2010).This paper presents the several behaviors of online users.The online environment has shifted dramatically in the last fifteen years, with orders of magnitude growth in both user content, as well as significant expansion of the capabilities of user's expect. Every day, new websites are emerge seeking to transform online paradigms and users with new types of offerings.

Social networking sites are skyrocketing in popularity and levels of user engagement, while traditional online communications paradigms like email is continue to see significant usage. Search over pages, listings, and multimedia is increasing in usage as well as complexity of resultsets.

Baum-Welch Style EM Approachon Simple Bayesian Models forWeb Data Annotationapproach was proposed by FatihGelgi and HasanDavulcu (2010).In this paper, weakly annotated data (WAD) are focused , and this is typically produced by a (semi) computerized information abstraction system from theWeb documents.By using statistical models, the extracted information has a certain level of accuracy and it can be exceededthat hasa capable of contextual reasoning such as Bayesian models. To re-annotate WAD,the contribution is the EM algorithm that operates on simple Bayesian models. EM estimates the parameters, i.e., by interacting Bayesian model on the given Web data, the prior and conditional probabilities. Bayesian classifier is skilled from current annotations at the expectation step,

III. PROBLEM STATEMENT

Most people in collaborative environments would be happy to share experiences with and give suggestions to others on specific problems. However, finding a right person is challenging due to the variety of information needs. This paper investigates how to enable such knowledge sharing mechanism by analyzing user data. Our goal is to find proper "advisors" who are most likely possessing the desired piece of fine-grained knowledge based on their web surfing activities.

For analyze the knowledge acquired by web users, they propose to log and analyze users' web surfing data (not only search, but also browsing activities, which reveal a user's knowledge gaining process). Users' interactions with the web can be segmented into different "tasks". Textual contents of a task are usually cohesive. We define a session as an aggregation of consecutively browsed web contents of a user that belong to the same task. People usually learn basic concepts first and then move towards advanced topics. 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. When pursuing a task, a user could

spend many sessions on a micro-aspect. Mining these micro-aspects (micro-knowledge) is critical: it can provide a detailed description of the knowledge gained by a person, which is the basis for advisor search.

IV. PROPOSED WORK

In this project,the existing Hidden Markov Model has some drawbacks.That are, HMM consumes larger memory and execution time is more. And also it provides less accuracy results.To Overcome these disadvantages Baum-Welch algorithm is proposed. This algorithm is the extension of Hidden Markov Model. This algorithm uses two matrices for finding expert in particular topic. That matrices are emission and transition matrix.Emission matrix finds the probability values at every hidden and observed State,that is at user and their related query. Transition matrix finds the probability values between two Hidden states,that is between users.After finding the transition and emission matrices values,the initial values are assigned to each user and set of observed sequences are taken.

Using this initial and matrix values,new transition and emission matrix values are calculated.Based on these values expert is selected in particular query.The aim of Expert search is extracting people who have more knowledge on the given query topic.This work is useful for company,which needs to know the users name who are expert in particular topic.Proposed algorithm more accurately finds the expert in particular query in company's database.

A. Architecture Diagram

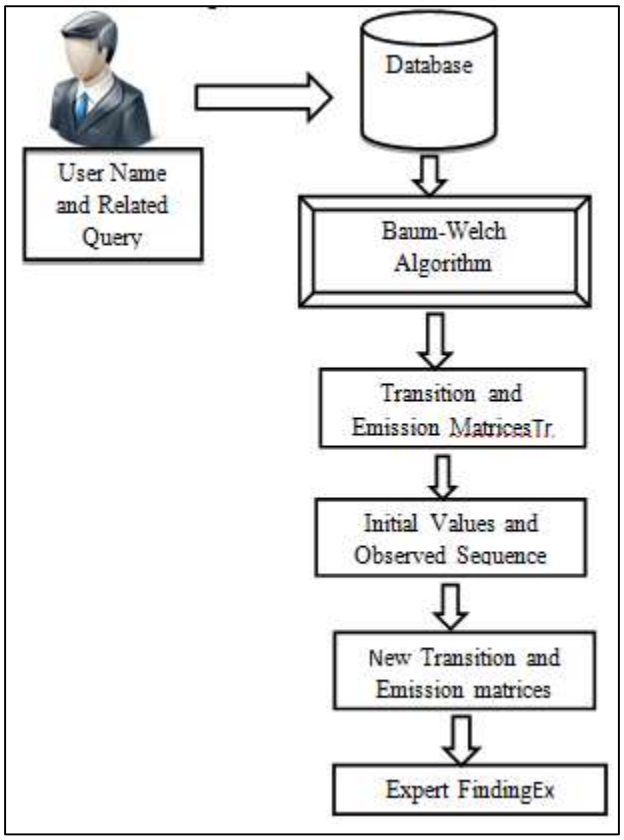

Fig.1:Architecture diagram

V. EXPERIMENTAL RESULTS

The Baum-Welch Algorithm is implemented and experimental result is obtained in the form of graph.By studying the graph it is evident that the implementation.

A. Accuracy

The Existing Hidden Markov Model gives the less accuracy results for finding the expert in particular Query.But when the Baum-Welch algorithm is implemented on query dataset it gives the accurate result in terms of finding the expert in particular topic.The Graph is plotted with the existing and proposed algorithm in the x-axis and accuracy percentage in y-axis.It can be noted that the accuracy has slightly increased when Baum-Welch algorithm is implemented.

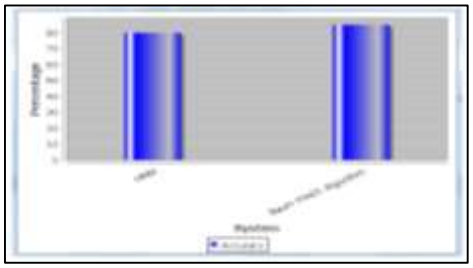

Fig. 2:Accuracy of BW Algorithm

B. Execution Time

This Graph Shows the execution time of BW algorithm which takes less milliseconds for execution than the HMM.Because before applying the HMM on query LaplacianEigenmap Gaussian DirichletProcess(LEGDP) should be used for Clustering websurfing data into tasks.The Graph is plotted with the algorithms in the x-axis and execution time in milliseconds in the y-axis.

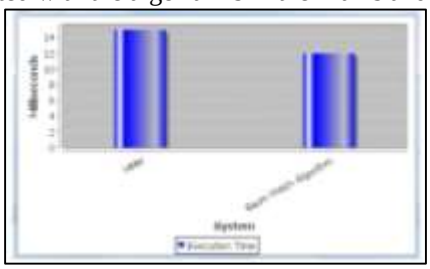

Fig. 3:Execution time of BW Algorithm

VI. CONCLUSION

Fine-grained knowledge reflected by people's interactions with the outside world is the key to solving this problem. Two step framework is used for extracting fine-grained knowledge and it is integrated with the classic expert search method for finding right advisors.For mining micro aspects we used discriminative-infinite Hidden Markov Model which is more costly. And also it provides less accuracy.It consumes large memory and longer execution time. We proposed Baum-Welch algorithm, which is the extension of HMM, for mining micro-aspects. This algorithm gives more accuracy than HMM. This algorithm use two matrices for finding probability which implies the best Advisor related to our query.Emission is at every hidden and observed State. Transition is between two Hidden states. In this work, we demonstrate the feasibility of mining micro-aspects of task for solving this knowledge sharing problem.

REFERENCES

[1] Ziyu Guan, Shengqi Yang, Huan Sun, MudhakarSrivatsa, and Xifeng Yan,2015,'Fine-Grained Knowledge Sharing in Collaborative Environments',vol. 27, no. 8
[2] Pushpalatha A M and Nirmal Kumar A, 2015,' A Survey on Using Baum-Welch Algorithm For Sharing Fine Grained Knowledge in Mutual Environments', Journal Of Applied Sciences Research, pp. 27-31
[3] KrisztianBalog, Leif Azzopardi, Maarten de Rijke, 2006,'Formal Models for Expert Finding in Enterprise Corpora,', in Proc. 29thAnnu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 43–50,
[4] D. M. Blei, M. I. Jordan and A. Y. Ng, 2003, 'Latent Dirichlet Allocation,', J. Mach. Learn. Res., vol. 3, pp. 993–1022.
[5] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen,2001, 'The infinite hidden Markov model,' in Proc. Adv. Neural Inf. Process.Syst.,pp.577–584.
[6] D. Blei and M. Jordan, 2006,'Variational inference for Dirichlet process mixtures,' Bayesian Anal., vol. 1, no. 1, pp. 121–143.
[7] Y. Fang, L. Si, and A. P. Mathur, 2010,'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, pp. 683–690.
[8] A. K. Jain,2010, 'Data clustering: 50 years beyond k-means,' Pattern Recog. Lett., vol. 31, no. 8, pp. 651–666.
[9] Y. Zhao, G. Karypis, and U. Fayyad, 2005, 'Hierarchical clustering algorithms for document datasets,' Data Mining Knowl. Discovery, vol.10,no.2,pp.141–168.
[10] R. Jones and K. Klinkner,2008, 'Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs,' in Proc. 17th ACM Conf. Inf. Knowl. Manage., pp. 699–708.
[11] R. Kumar and A. Tomkins,2010, 'A characterization of online browsing behavior,' in Proc. 19th Int. Conf. World Wide Web, pp.561–570.
[12] H. Deng, I. King, and M. R. Lyu,2009, 'Formal models for expert finding on DBLP bibliography data,' in Proc. IEEE 8th Int. Conf.Data Mining, pp. 163–172.
[13] M. Belkin and P. Niyogi,2001, 'LaplacianEigenmaps and spectral techniques for embedding and clustering,' in Proc. Adv. Neural Inf. Process. Syst., pp. 585–591.
[14] P. R. Carlile, 1998,'Working knowledge: How organizations manage what they know, Human Resource Planning,' vol. 21, no.4,pp.58-60
[15] H. Deng, I. King, and M. R. Lyu,2009, 'Formal models for expert finding on DBLP bibliography data,'in Proc. IEEE 8th Int. Conf. Data Mining, pp. 163–172.
[16] J. Liu and N. Belkin, 2010,'Personalizing information retrieval for multi-session tasks: The roles of task stage and task type,' in Proc. 34th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf.Retrieval,pp.26-33

[17] X. Liu, W. B. Croft, and M. Koll,2005, 'Finding experts in community based question-answering services,' in Proc. 14th ACM Int. Conf. Inf. Knowl. Manage., pp. 315–316.
[18] C. Rasmussen,2000, 'The infinite Gaussian mixture model,' in Proc. Adv. Neural Inf. Process. Syst., pp. 554–560.
[19] P. Serdyukov, H. Rode, and D. Hiemstra, 2008,'Modeling multi-step relevance propagation for expert finding,'in Proc. 17th ACM Conf. Inf. Knowl. Manage., pp. 1133–1142.