Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 7, July -2017

Fine Grained Knowledge Sharing Mechanism with Hierarchical Structure Using Iterative d-iHMM

Saraswati Sonkale[1], Mrs. V.L.Kolhe[2]

[1] *Department of Computer Engineering, D.Y.Patil COE, Akurdi,Savitribai Phule Pune University, Pune*
[2] *Department of Computer Engineering, D.Y.Patil COE, Akurdi,Savitribai Phul/e Pune University, Pune*

Abstract — *In collaborative environments, individuals may attempt to acquire similar information on the web keeping in mind the end goal to pick up data in one domain. For instance, in an organization a few divisions might progressively need to purchase business insight software and representatives from these offices may have to learn online about different business intelligence tools & their characteristics independently. A two-stage framework is used for mining fine-grained data:(1) Web surfing information is grouped into clusters by using hierarchical clustering; (2) a novel discriminative infinite Hidden Markov Model is used iteratively to mine fine-grained data from every task. The excellent master inquiry technique is connected to mined results to discover appropriate individuals for information sharing.*

Keywords- *Advisor search, text mining, Dirichlet processes, graphical models, Hierarchical Clustering, d-iHMM.*

I. INTRODUCTION

Interact with the web and with partners or companions to obtain data is a day by day routine of numerous people. In a community situation, it could be basic that individuals attempt to procure same data on the web keeping in mind the end goal to increase particular information in one area. For case, in an organization a few divisions might progressively need to purchase business intelligence (BI) applications, and representatives from these divisions may have study online about diverse of information about BI tools and their elements. In an examination lab, individuals are regularly concentrated around tasks which require comparable basic information. An analyst might need to tackle an information mining issue by utilizing nonparametric graphical models for which they are not familiar with but have been concentrated by another analyst some time recently. In these cases, depending on a correct individual could be much more productive than studying without anyone else's input since individuals can give processed data experiences and live associations as contrasted with the web.

For the first situation, it is more profitable for a worker to get advices on the decisions of BI applications and clarifications of their components from experienced representatives; for the second situation, the first analyst could get proposals on model configuration and study materials from the second scientist. Many people in synergistic situations would be glad to offer recommendations to others on particular issues. Discovering a perfect individual is a great job because of the assortment of data needs. So knowledge sharing mechanism will be investigated by analyzing user data.

A micro-aspect could be roughly defined as a significantly more cohesive subset of sessions in a task. For example, the task learning Java might contain Java IO and Java multithreading as two micro-aspects. 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. The goal is to detect peoples online learning activities (e.g. learning Java IO) in session data reflected by subsets of sessions rather than discerning topics hierarchically in sessions (e.g. Java with IO as its subtopic). Mining the semantic structures in sessions is important for advisor search.

A two-step framework for mining fine grained knowledge (micro-aspects) is developed: At first stage, the tasks are made from sessions. Hierarchical clustering is used for clustering the lots of web sessions. It is useful for creating partitions. A non parametric scheme is adapted because number of tasks is difficult to predict. At second stage extract micro-aspects from task and evolution patterns in each task. A background model is introduced iterative d-iHMM in order to enhance the discriminative power. Finally, language model based expert search method is applied over the mined micro-aspects for advisor search.

II. RELATED WORK

The aim of the expert search is recapturing users who have best knowledge on the given inquiry. A knowledge base involves early approaches which contains the detail information of expert knowledge of the peoples within an organization.

Expert search

An area of analysis like finding expert became a popular when TREC enterprise track started in 2005[5]. Balog et al. constructed a model of language framework for expert search. It describes an approach of document-centric in which the related documents to a inquiry is first searched and then acquire for every users. Score is displayed with Candidate for relevance of documents. By this process a probabilistic model was prepared in a generative model. Balog et al. showed this Model gives performance well and it became one of the most outstanding methods for searching expert. Other methods have been proposed for enterprise expert search but the nature of these methods.

Analysis of Search Tasks

Recently, query logs are used by researchers to have concentrate on finding, designing and observing users search tasks. Jones and Klinkner found that tasks which are to be searched are inter-weaved and these are used classifiers to partition the number of user queries in sequence into the tasks[9]. Using a stage three and type two controlled experiment, Alexandre and Megretski,Munther A. merged stage and type of the with dwell time to prove the efficiency of a result document[8]. Konstantinos Bousmalis modeled classifiers to find out similar-task questions for a given inquiry and to conclude whether a task will resume by user[12]. Ji et al. used regularization of graph to find out search tasks in query records. Wang et al. evaluated the cross-session for the mining problem of search task as a semi supervised clustering problem where the dependency structure related to the queries in a search task was absolutely designed and a set of automatic annotation rules were proposed as weak supervision. Jurgen Van Gael,Yunus Saatci introduced Beam Sampling for the iHMM. Beam sampling is an inference algorithm for the iHMM. Beam sampling also consist of slice sampling, so number of states are limits by using dynamic programming considered at each time step to a finite number. Using the beam sampler this presents applications of iHMM inference on change point detection and text prediction problems [2].

Hierarchical Clustering

Hierarchical clustering or Hierarchical Cluster Analysis (HCA) is a method of formulating clusters which is used to explore a hierarchy of clusters. Two approaches are there for hierarchical clustering[6].

Agglomerative : This is a bottom up approach in this each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

Divisive: This is a top down approach in which all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Greedy method is used to determine the merges and splits. Dendrogram are used to show the results of hierarchical clustering. The complexity of Agglomerative clustering is $O(n^2 log(n))$ which makes algorithm too slow for huge set of data. The complexity Divisive clustering with an exhaustive search is $O(2^n)$ which is also worse.

Padhraic Smyth introduces a probabilistic model-based approach for clustering *sequences*, using hidden Markov models (HMMs), to solve the following two issues which are addressed. First, a novel parameter initialization procedure and second the more difficult problem of determining the number of clusters *K*, from the data is investigated[10].

Emanuele Coviello, Antoni B. Chan, Gert R.G. Lanckriet developed a novel algorithm to cluster HMMs based on the hierarchical EM (HEM) algorithm. This algorithm i) clusters a given collection of HMMs into groups of HMMs that are similar, in terms of the distributions they represent and ii) characterizes each group by a "cluster center", it's a novel HMM that representative for the group[7].

Manuele Bicego1, Vittorio Murino introduced a novel scheme for HMM based sequential data clustering using similarity based paradigm recently introduced in the supervised learning context. Using this approach a new representation space is built in which each object is described by the vector of its similarities with respect to a pre-determinate set of other objects. These similarities are determined using hidden Markov models. Clustering is then performed in such a space. By way of this the difficult problem of clustering of sequences is thus transposed to a more manageable format[5].

Iterative d-iHMM

Frank Wood introduced A Discriminative Nonparametric Bayesian Model. On many existing graphical models with latent variables, a Nonparametric methods have been successfully applied. Optimal number of hidden states are automatically learns by this model with the given a specific dataset. This is achieved by using Hierarchical Dirichlet Processes (HDP)[13].

III. SYSTEM ARCHITECTURE

A framework having two steps is proposed for extracting expert knowledge(micro-aspects): In the first step, sessions are used to formulate the tasks . To cluster the sessions Hierarchical clustering based on Dirichlet Process is designed (DP).Then mining micro-aspects from sessions iteratively from each task. The challenges are: the number of micro-aspects in a task are unknown; for different micro-aspects sessions of a task are textually similar micro-aspect of the sessions might not be consecutive. To mine micro-aspects in each task iteratively an Iterative discriminative infinite

Hidden Markov Model (d-iHMM) is developed. Finally for advisor search applied a language model based expert search method over the mined micro aspects.

Sessions: Consecutively browsed web contents are aggregated in a session of a user that part of the same concepts. In this analysis sessions are atomic units. The content within sessions in a task can increase gradually: Basic concepts are learn by the peoples first then they move towards gaining in depth knowledge.

Micro-aspect: A task can be further divided into fine-grained things (called micro-aspects). Micro-Aspects are referred as significantly more huge subset of sessions in a task. For example, the learning Java task might contain Java IO and Java multithreading as two subset. When pursuing a task a user could spend many sessions on a micro aspect. It's difficult to extract these micro-aspects (micro-knowledge).

Hierarchical Clustering: For data clustering Hierarchical Clustering used. The cluster are fitted by assigning query data points to the multivariate normal components that maximize the component posterior probability given the data.
In this N no. of items are clustered and an N*N distance (or similarity) matrix is drawn. This kind of hierarchical clustering is called *Agglomerative* because it merges clusters iteratively.

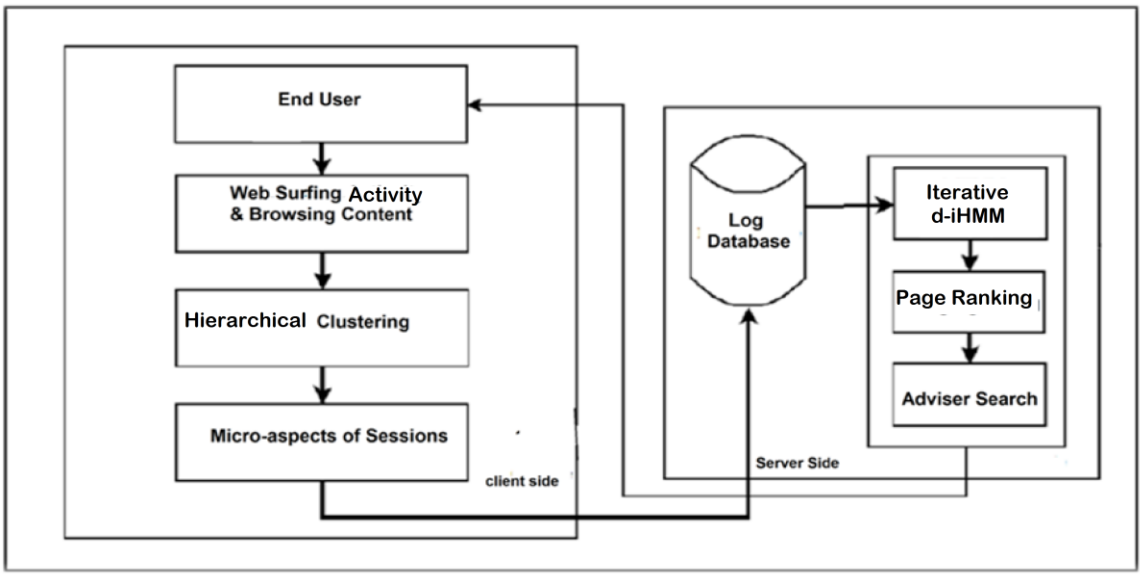

Figure 1:- System Architecture of Fine Grained Knowledge Sharing Framework

Algorithm:

The system implements following algorithm for recommending the expert person:-

Algorithm: Apply Iterative d-iHMM model to retrieve the micro-aspect through hierarchical data.
Input : Users web surf database, queries, clusters
Output : Recommendation of Expert person

1. Enter the query.
2. Randomly all the session are searched.
3. Sessions are clustered using Hierarchical clustering method.
4. Mining over clustered sessions by applying d-iHMM model iteratively to retrieve the hierarchical data.
5. Apply the adviser search for ranking of the candidate.
6. Recommend the expert person with his sessions.

IV. RESULTS

From real collaborative environments the dataset is collected as a web surfing data. The sessions surfed by the users is collected as dataset which is of one month. The sessions are segmented using sequence of http contents of each web users by considering the following rule: between two consecutive contents place a session boundary if the difference between timestamps should be minimum 10 minutes from each other Dataset contains nearly 500 sessions.

Whenever any query is entered for searching information fine grained knowledge sharing mechanism first shows all the sessions regarding to the query. Some of that sessions may be generally related or Truly related. Following result shows the output for the query: "Java Inheritance".

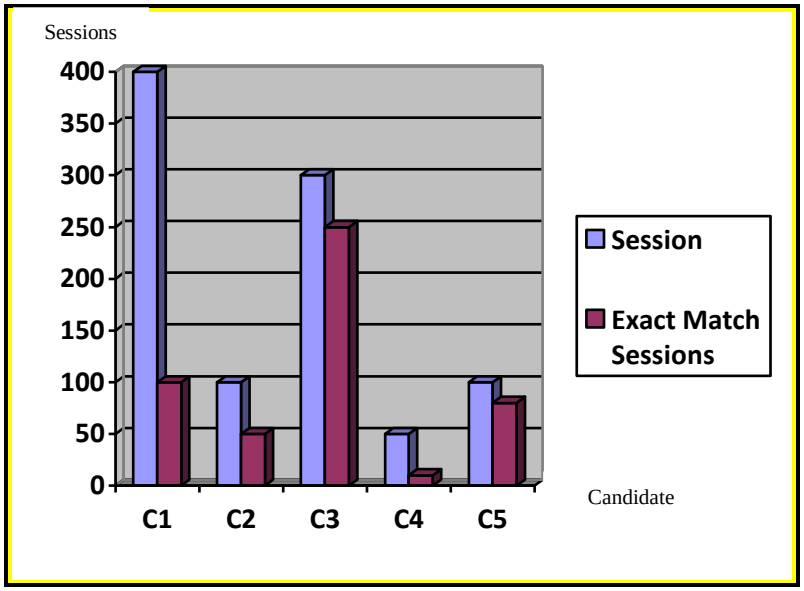

Figure 2: Sessions using Micro-Aspect-Based Scheme

Table1: Sessions using Micro-Aspect-Based Scheme

Candidate	Exact match Session	Total Session
C1	100	400
C2	50	100
C3	250	300
C4	20	50
C5	80	100

The result in figure: 2 states that for the query, the micro aspect- based scheme ranks C1 and C3 at the top while the session- based scheme ranks C1 and C3 at the top respectively. Candidate C1 has generated 100 sessions and from that 400 sessions are truly related. Candidate C3 has generated 300 sessions and from that 250 sessions are truly related, so candidate C3 have fine grained information about query "Java Inheritance" than all of other candidates.

For evaluation of clustering methods three popular evaluation metrics are used: Purity, F-measure and Normalized Mutual Information (NMI).

Purity: It tries to map each cluster to the class in the ground truth which is the most frequent in the cluster. It is defined as follows:

$$\text{Purity}(\Omega,\Xi) = 1/n \sum_{k=1}^{K} \max_{j \in \{1 \ldots J\}} |w_k \cap \mathfrak{C}_j|$$

Where, $\Omega = \{w_1 \ldots w_K\}$ is the set of clusters and $\Xi = \{\mathfrak{C}_1 \ldots \mathfrak{C}_j\}$ is the set of ground truth classes.

Normalized Mutual Information (NMI): F-measure operates on pairs of objects (i.e. sessions). The clustering result is treated as n(n-1)/2 decisions each corresponds to a pair of objects and decides whether they belong to the same class.

Depending on this the precision and recall scores are computed and the F-measure is computed as the harmonic mean of precision and recall. NMI is defined as:

$$NMI\ (\Omega,\Xi)= I(\Omega, \Xi)/\ (1/2(H(\Omega)+H(\Xi)))$$

Where I(.) and H(.) represent mutual information and entropy respectively.

Table 2: Performance Comparison of Two Different Clustering Methods

Methods	Purity	NMI
Hierarchical Clustering(HC)	.700	.775
LEGDP	.667	.698

Table:2 shows the results of the clustering performance measured by Purity and NMI. LEGDP approximately captures the number of tasks. It also shows the number of clusters learned by Hierarchical Clustering(HC) which is outperforms. LEGDP and HC can achieve comparable performance.

VI. CONCLUSION and FUTURE SCOPE

Knowledge sharing mechanism is a new technique for finding expert in collaborative environment. Finding the right person in an organization with the proper skills and knowledge is often crucial to the success of projects. To find right person, developed an iterative novel discriminative iHMM to extract micro aspects with hierarchical data. This method of finding expert is better than traditional expert search problem where expert search goal is to find domain experts based on associated documents. Probes genuine web surfing data appeared empowering results. This could iteratively apply d-iHMM on the scholarly micro-aspects to determine a chain of command. The fundamental inquiry model is refined. A two-step framework to mine fine-grained knowledge is implemented and it integrated with the page ranking algorithm for finding right advisors.

Currently all the web sessions are recorded in the database, no matter it may be personal search or official search. So privacy can become an issue. If privacy is maintained it will not record the session which are personal to an employee. So leaved this possible improvement for the future enhancement.

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, 2015
[2] Jurgen Vanv Gael, Yee Whye Teh, Zoubin Ghahramani," Beam Sampling for the infinite Hidden Markov Model ", In proceeding of the International Conference on Machine Learninig, 2010.
[3] Georgios Kotsalis, Alexandre Megretski, Munther A. Dahleh, "A Model Reduction Algorithm for Hidden Markov Models ", In Proceedings of the 45th IEEE Conference on Decision and Control Manchester Grand Hyatt hotel San Diego, CA, USA, December 13-15, 2006.
[4] Nick Craswell MSR Cambridge, "Overview of the TREC- 2005 Enterprise Model ",In Proceedings of the Text Retrieval Conference Gaithersburg, MD 2005.
[5] Krisztian Balog, Ian Soboro,"Overview of the TREC 2008 Enterprise Track ", University of Amsterdam, NIST, Text REtrieval Conference Gaithersburg, MD, 2008.
[6] Jingxuan Li, Bo Shao, Tao Li, and Mitsunori Ogihara, Member IEEE, "Herarchical Co-Clustering: A New Way to Organize the Data ", IEEE Transaction on Multimedia, Vol 14, no. 2. April 2012.
[7] Ana L. N. Fred and Anil K. Jain, "Combining Multiple Clustering Using Evidence Accumulation ", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 6, June 2005.
[8] Alexandre Megretski,Munther A Dahleh, "A Model Reduction Algorithm for Hidden Markov Models "Georgios Kotsalis, In Proceedings of the 45th IEEE Conference on Decision and Control 2006.
[9] Ujjwal Maulik, Sanghamitra Bandyopadhyay, Indrajit Saha, "Intgrating Clustering and Supervised Learning for Categorical Data Analysis ", IEEE Transactions on Systems and Humans, July 2010.
[10] M. Srinivas and C. Krishna ,Mohan, "Efficient Clustering Approach using Incremental and Hierarchical Clustering methods ", IEEE 2010.
[11] Ana L.N, Anil K. Jain, "Combining Multiple Clusterings Using Evidenc Accumulation ", IEEE Transactions on Pattern Analysis and Machine Intelligence, June 2005.

[12] Konstantinos Bousmalis,Louis Philippe Morency,Stefanos Zafeiriou,Maja Pantic, "A Discriminative Nonparametric Bayesian Model: Infinite Hidden Conditional Random Fields ".
[13] Frank Wood, "Hidden Markov models: from the beginning to the state of the art ", Columbia University, November, 2011 Journal of Computer Science and Technology July 2010.
[14] Dilan Gorura,Carl Edward Rasmussen, "Dirichlet Process Gaussion Mixture Models:choice of the Base Distribution ", IEEE Transaction on System and Cybernetics-part a: Systems and Humans,vol. 40, no. 4, July 2010.
[15] Cody Hudson, Benard Chen and Dongsheng Che, "Hierarchically Clustered HMM for protein sequence extraction with variable length ", IEEE Transaction on pattern Analysis and Machine Intelligence,August 2014 .
[16] Trushagni Bhoi, Sayali Shinde, Kajal Kajale,Prof. NitinShivale, "Text Mining using Hybrid Algorithm ", International Engg Research Journal (IERJ) vol. 1, issue 9, page 815-818, 2015 ISSN 2395-1621

AUTHORS

Saraswati Bhagwan Sonkale, Pursuing M.E. in Computer Engineering at D. Y. Patil *College of Engg. Akurdi, Pune.*

