Effective Online Knowledge Sharing using Refined Search

Atanu Samadder[1], Mandar Kale[2], Shivender Singh[3], Sayali Jojare[4]
[1, 2, 3, 4] JSPM's Imperial College Of Engineering, Wagholi, Pune, India

Abstract- *Present environment is a synergistic and people tend to search relevant information on a specific topic online, and hence in that joint many familiar searches is bound to happen. For example: many people want to learn programming language, assuming the language is JAVA, If a person searches tutorials on JAVA, he will get the tutorial results along with other people who also has have searched the related topic and have done some research on it. So, the user can either continue searching or contact the person who has done some research on that particular topic. We use refined online Knowledge sharing in joint environment. Two step frame work is used in this search-(1)web surfing cluster is made.(2) Hidden Markov Model is developed to mine fine-grained aspects in each task. When it is integrated with expert search, the searching accuracy improves successively, in comparison with traditional search.*

Keywords- Advisor Search; Markov Model; Fine-grain; Expert Search.

I. INTRODUCTION

Present Generation has a shared environment, hence it is easy to inquest any topic or any data which is required. Human beings are social animals hence learning from other people is more salutary. For example if a person searches a topic online, that topic might be searched or explored by many other people. The learner can always enjoin other people for better understanding of that particular topic. In the shared environment people willing to share their knowledge are most welcome, But the tricky part is to find appropriate person with substantial data. In this paper, we examine how to enable such knowledge sharing mechanism by analysing user information.

The contributions of this work are summarized as follows. (1) We introduce the fine-grained knowledge sharing problem in joint environments. The motive is not only to find domain experts but a person who has the desired a specific knowledge on a topic. The problem is significant in practice by learning from a confidant (if she/he is easy to find) it might be more efficient than studying on the web (though not always).

An illustrative example is shown in Fig. 1. One can use "tcpdump" to intercept a sequence of web surfing activities , IP packets for each member. The scene is, Rahul starts to surf the web and wants to learn how to develop a Java multithreading program, which has already been studied by Atanu (red rectangle). In this context, it might be a good idea to consult Atanu, rather than studying by herself. We aim to provide such commendations by analysing the surfing activities automatically. In this example, not necessarily Atanu is an expert in every facet of Java programming; however, due to his frequent surfing activities in Java multithreading, it is acceptable to assume that he has gained adequate knowledge in this area so that he can help Rahul (in real life implementation we could set a threshold on the amount of

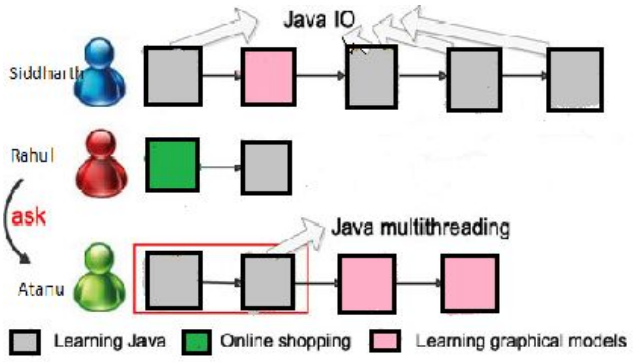


Fig. 1. An illustrative toy example for knowledge sharing in a collaborative environment.

related surfing information to test significance). Even if Atanu is still learning, he could share his experiences in learning and possibly suggest good learning materials to Rahul, thus saving Rahul's effort and time.

II. RELATED WORK

In this section we review and analyse some research fields that are related to our work: expert search, analysis of user search function and topic modelling.

2.1 Expert Search

Expert search aims at retrieving people who have competence on the given query topic. Previous approaches involved building a knowledge base which contains the descriptions of people's skills within an organization. Expert search became a talk about research area since the start of the

TREC enterprise in 2005. Balog et al. advised a language model framework for expert search. The Model 2 is a document-centric approach which first enumerate the relevance of documents to a query and accumulates for each candidate the relevance scores of the documents that are associated with the candidate. This process was developed in a generative probabilistic model. Model 2 proved to be better and it became one of the most extrusive methods for expert search.

2.2 Analysis of User Search Function

In recent times, researchers have complete focus on detecting, analysing and modelling user search functions from query logs. Here we name some representative work of the users. Raj and Rahul found that search tasks or functions are interweaved and used classifiers to segment the sequence of user queries into respective tasks. Prakash and parul combined the task stage and task type with abide time to predict the advantage of a result document, using a two-type and three-stage controlled experiment. Ganesh used graph regularization to recognize search functions in the query logs . Pranav designed classifiers to recognize same-task queries for a given query and to anticipate whether a user will continue a task. Helen developed the cross-session to mine search function problem as a semi-supervised clustering problem where the dependency structure among the queries was explicitly modelled and a set of automatic annotation rules were proposed as low supervision. This research tries to recover tasks from user's search behaviours' and bears some similarity to our work. Nevertheless, our work differs from the following aspects. First, we use the general web surfing contents, including search, not the search engine query logs. Query logs does not support subsequent surfing activity after the user clicked a relevant result online. Moreover, it is being estimated that 50 percent of a user's online page views are content browsing data. Web surfing data provides more precise information about the knowledge retrieving. Although different methods were proposed for mining search functions in query logs, these methods cannot be implemented in our pre-set setting since they extract query log specific properties. Second, none of the above tried to mine fine data aspects for each task. People could spend some effort on one fine grained aspect of a function and generate many contents when studying. In brief, fine-grained aspects can provide a good description of the knowledge gained by the user. Finally, none of existing implementations analyse user online behaviours' which is not limited to search behaviours'.

2.3 Topic Modelling

Topic modelling is a tool for analysing topics in any document collection and he most prevalent topic modelling method is Latent Dirichlet Allocation (LDA). Based on LDA, many topic modelling methods have been advised, For e.g. the dynamic topic model for any sequential data and the hierarchical topic model for building topic hierarchies. The Hierarchical DP (HDP) model is the nonparametric version of LDA. However, The problem is not a topic modelling. Our goal is to gather the semantic structures of people's online knowledge gaining activities from their web surfing, i.e. identifying groups of sessions which represent tasks for e.g. learning "Java" and micro-aspects for e.g. learning "Java multithreading". Topic modelling decomposes a document into topics but after applying topic modelling methods on the session data, it is still difficult to find the right advisor. This is because a person can have many sessions containing partially relevant topics which then will be ranked unexpectedly high, due to this accumulation of relevance among sessions, grouping it into micro-aspects can help for finding the right advisor.

III. CLUSTERING SESSIONS

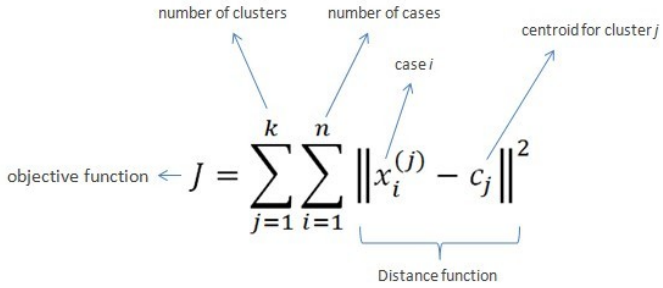
3.1 What Do You Mean By Clustering?

1. A cluster is a subset of data which are similar. Clustering (also called unsupervised learning) is the practice of dividing a dataset into groups such that the members of each group are as similar(close) as likely to one another, and different groups are as dissimilar (far) as possible from one another.
2. There are many applications for cluster study. For example, in business, cluster analysis can be used to discover and exemplify buyer segments for marketing purposes and in biology.

3.2 Partition Method

1. K-Means clustering intends to separation n objects into k clusters in which each item belong to the cluster with the nearest mean. This method produces accurately k different clusters of greatest possible distinction.

2. The best number of clusters k leading to the greatest partition (space) is not known as a priori and must be computed from the data. The objective of K-Means clustering is to diminish total intra-cluster variance, or, the squared error purpose:



$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

3. K-Means is relatively an efficient method. However, we need to stipulate the number of clusters, in progress and the final results are receptive to initialization and often terminates at a local optimum. Unluckily there is no global theoretical technique to find the optimal number of clusters. A practical approach is to contrast the outcomes of multiple runs with different k and select the best one based on a predefined measure. In general, a large k possibly decreases the error but increases the risk of over fitting.

4. An important problem in clustering is how to determine the likeness between two objects, so that clusters can be formed from items with high match within clusters and low match between clusters. Commonly, to compute similarity or dissimilarity between objects, a distance evaluate such as Euclidean, Manhattan and Minkowski is use. A space function returns a lower value for pairs of items that are more like to one another.

5. Following formulas are used to find out centroid and matrix;

Euclidean $\sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k} |x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k} (|x_i - y_i|)^q \right)^{1/q}$

Fig 2: List of formula's

1. Clustering by Gaussian Mixture Model

When using probabilistic models for making clusters, mostly we use Gaussian mixture model and it is probabilistic version of k-means. However, for applying Gaussian distributions in our case, the data dimensionality D0 is very high mostly above 10k. Therefore, first we apply the well-known Laplacian Eigen map (LE) technique to reduce the dimensionality of the data from D0 to D where D0 _ D. We choose LE since it could also capture the nonlinear complex structure of a task, e.g. the topics could evolve and move a bit, which could be characterized by the "half-moon" structure

IV. USE OF HIDDEN MARKOV MODEL

Hidden Markov models (HMMs) are one of the most used and popular methods for machine learning and statistics for modelling sequences for speeches. An HMM defines a probability distribution over sequences of observations (symbols) y = {y1,….,yt,….,yr} by invoking another sequence of unnoticed, or hidden, discrete state variables s = {s1,…,st.....sr} The main idea of HMM is that the sequence of hidden states has Markov dynamics—i.e. given st, sr is independent of sρ for all r < t < ρ —and that the observations yt are independent of all other variables given st. The model is defined in terms of two sets of parameters, the transition matrix whose ijth element is P(st+1 = jjst = i) and the emission matrix whose iqth element is P(yt = qjst = i). The usual procedure for estimating the parameters of an HMM is the Baum-Welch algorithm, a special case of EM, which estimates required values of two matrices n and m corresponding to the number of transitions and emissions respectively, where the expected value is taken over the posterior probability of hidden state sequences . Both the standard estimation procedure and the model for HMMs suffer from important limitations. First, likelihood estimation procedures do not consider the complexity of the model, making it hard to avoid over and under fitting. Second, the model structure has to be specified in advance. Motivated in part by these problems we see that there have been some attempts to approximate a full Bayesian analysis of hidden markov models which integrates the parameters, rather than optimisation. It has been to approximate such Bayesian integration both using variation methods and by accustomed on a single most likely the hidden state sequence.

V. EXPERT SEARCHING

Addressing a problem with a approach of identifying expertise within a wide range of organization has lead to the development of many category of search engines. In Today's search engine category it is been known for a quite time called as Expert Search. McDonald and Ackerman distinguish several aspects of expert search results, including the expertise identification ("Who are the right people who has done research on this Topic?") and expertise selection ("What

does person "A" knows about the particular topic?").Therefore, In this we are focused on these questions. Early expert search used a database containing a description of peoples' skill sets within the organization. The static nature of the databases can sometime renders them as antiquated and incomplete. Moreover, expert search queries tend to be fine-grained and specific, but we observed that the descriptions of expertise tend to be generic . To address these disadvantages a number of systems have been thought of which aimed at automatically discovering latest expertise information from secondary sources. Usually, this has been performed in some specific domains. For example, there have been attempts to use email interaction for expert search in discussion threads and personal mails. Campbell et al. analysed the link structure defined by authors and the mails delivered to the receiver is a modified version of the Hyperlink-Induced Topic Search (HITS) algorithm to identify authorities Another approach of using email communications focused on detecting communities of expertise, positing that the noticeable behaviour between individuals would indicate that the person has done some research in a specific area, again using the HITS algorithm .

VI. BUSINESS LOGIC ARCHITECTURE

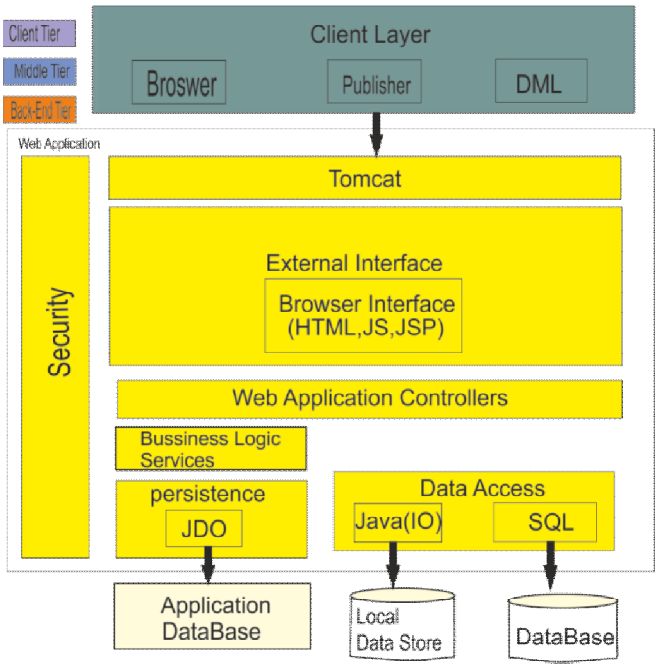


Fig 3. Business logic Architecture.

In the today's competitive market to improve the business processes to act accordingly with the increasing changes. The evolution of the business process normally leads to a change in the employed software systems. Software evolution is a lengthy and costly task , when the documentation of a system is lost, outdated or not available. In this paper a business-logic-based framework for evolving software systems is used. The goal is evolving software in a higher abstract layer.

VII. IMPLIMENTATION

After overcoming all the complex errors and web searches and data bases we were able to make the user friendly system that makes understanding any subject that the user wants to or interested to research on.

The following picture shows the Register Page of the system implementation og the knowledge sharing expert System.

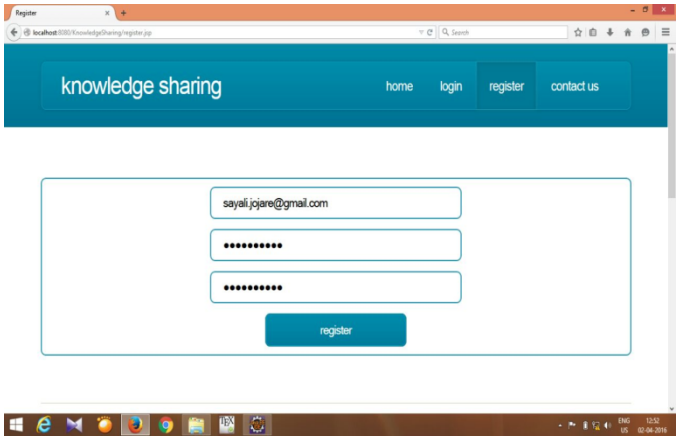


Fig 4.The register Page

The user has to register to access all the files he want to search. He is asked to enter his/her email id and record a password for authentication.

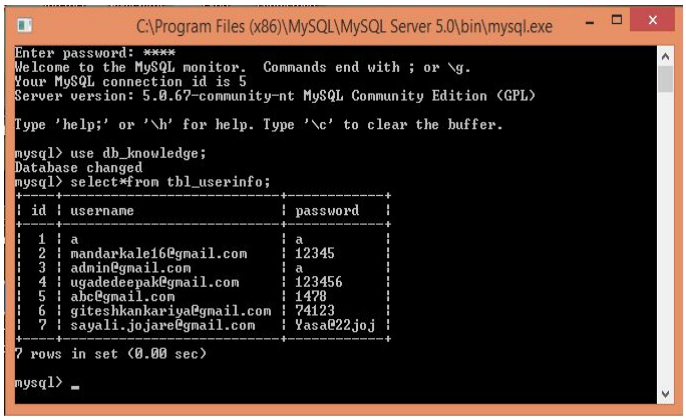


Fig 5.Database in MySQL

All the user information is stored in the database and the more he rates any document he finds interesting, his personal ratings too go higher. The database will contain user rating and other information necessary for the knowledge sharing and real time communication

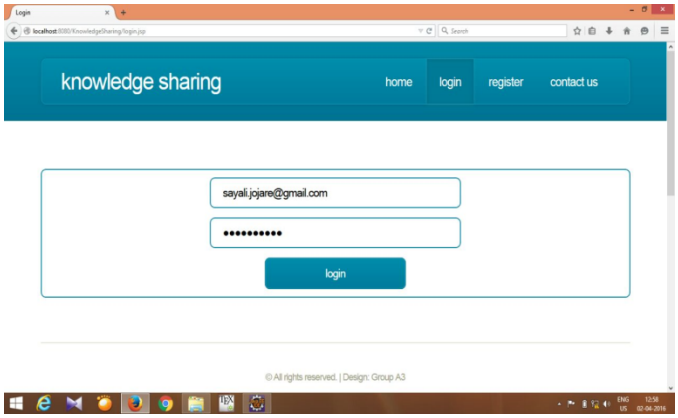


Fig 6.The Login Page

Once the user has registered he can now log in, the login page is shown above.

After the user logins, it can access hi/her search page or the front page, the front page contains preloaded documents, as this is demonstration project, the amount of documents are less, although the concepts is implemented without any error.

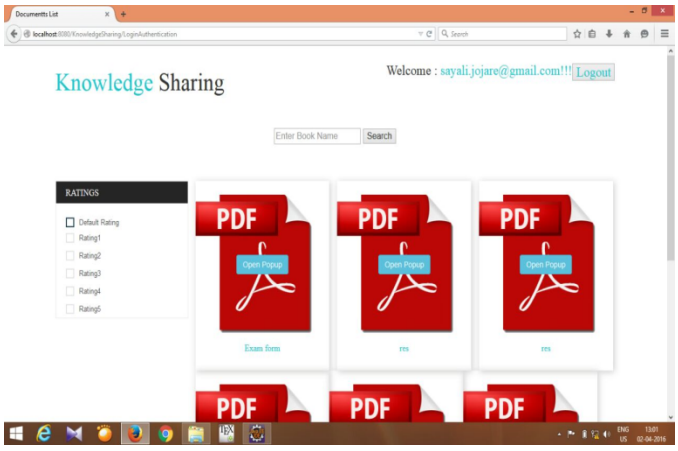


Fig 7.Front Page

The documents can be downloaded multiple times and every time the user downloads it he/she can rate the document according to the usefulness of that document.

The rating is updated every time the any user rates that document giving the average rating of the document and is placed according to that in the front page off any user web searching profile.

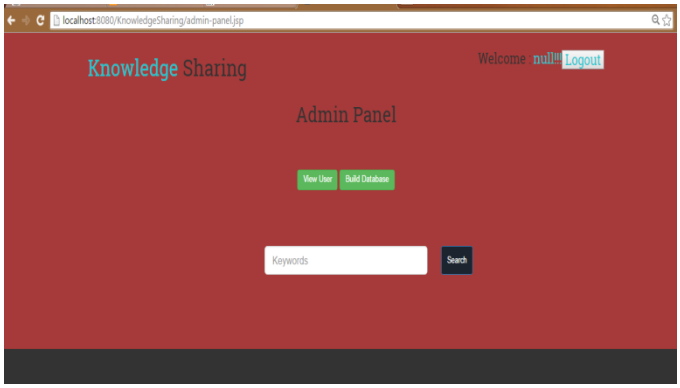


Fig 8.Admin Page

The files that are uploaded ,are uploaded through the admin page. The Admin page will contain the admin login and view user details like user name, password, file uploading option.

This page will be able to control all the user ratings and expert user priority .

VIII. CONCLUSIONS

We introduced fine-grained knowledge sharing in shared environments, which is desirable in practice. We identified fine-grained knowledge reflected by people's interactions with the outside world as the key to solving the problem. We introduced a two-step framework to mine valuable knowledge and integrated it with the current expert search method for finding the required advisors. Experiments on real web surfing showed an improved and encouraging results. We found some issues for this problem. (1) The fine-grained knowledge could have a hierarchical structure. For example, "Java IO" can inherit some files which act as a sub-knowledge file like "File IO" and "Network IO" . We could iteratively apply d-iHMM on the learned micro-aspects to derive a hierarchy, but how to search over this hierarchy is not a trivial problem. (2) The basic search model can be refined, e.g. incorporating the time factor since people gradually forget as time flows. (3) Privacy is also an issue. In this work, we demonstrate the feasibility of mining task micro-aspects for solving this knowledge sharing problem. We leave these possible improvements to future work.

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] M. Belkin and P. Niyogi, "Laplacian Eigenmaps and spectral techniques for embedding and clustering," in

Proc. Adv. Neural Inf. Process. Syst., 2001, pp. 585–591.

[3] D. Blei and M. Jordan, "Variational inference for Dirichlet process mixtures," Bayesian Anal., vol. 1, no. 1, pp. 121–143, 2006.

[4] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, "Hierarchical topic models and the nested Chinese restaurant process," in Proc. Adv. Neural Inf. Process. Syst., 2003, pp. 17–24.

[5] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in Proc. Int. Conf. Mach. Learn., 2006, pp. 113–120.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.

[7] P. R. Carlile, "Working knowledge: How organizations manage what they know," Human Resource Planning, vol. 21, no. 4, pp. 58–60, 1998.

[8] N. Craswell, A. P. de Vries, and I. Soboroff, "Overview of the TREC 2005 enterprise track," in Proc. 14th Text REtrieval Conf., 2005, pp. 199–205.

[9] H. Deng, I. King, and M. R. Lyu, "Formal models for expert finding

[10] 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.

[11] T. S. Ferguson, "A Bayesian analysis of some nonparametricproblems," Ann. Statist., vol. 1, no. 2, pp. 209–230, 1973.

[12] A. K. Jain, "Data clustering: 50 years beyond k-means," PatternRecog. Lett., vol. 31, no. 8, pp. 651–666, 2010.

[13] M. Ji, J. Yan, S. Gu, J. Han, X. He, W. Zhang, and Z. Chen,"Learning search tasks in queries and web pages via graphregularization," in Proc. 30th Annu. Int. ACM SIGIR Conf. Res.Develop. Inf. Retrieval, 2011, pp. 683–690.

[14] R. Jones and K. Klinkner, "Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs," in Proc. 17th ACM Conf. Inf. Knowl. Manage., 2008, pp. 699–708.

[15] 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.

AUTHOR PROFILE

Atanu Samadder is currently pursuing his Bachelor's degree in Computer science from JSPM's Imperial College Of engineering, Wagholi, Pune, Maharashtra, India.

Shivender Singh is currently pursuing his Bachelor's degree in Computer science from JSPM's Imperial College Of engineering, Wagholi, Pune, Maharashtra, India.

Sayali Jojare is currently pursuing her Bachelor's degree in Computer science from JSPM's Imperial College Of engineering, Wagholi, Pune, Maharashtra, India.

Mandar Kale is currently pursuing his Bachelor's degree in Computer science from JSPM's Imperial College Of engineering, Wagholi, Pune, Maharashtra, India.