A REVIEW ON KNOWLEDGE SHARING IN COLLABORATIVE ENVIRONMENT

[1]Manasi Khot, [2]Bhagyashree Wani, [3]Payal Shah, [4]Sonali Raut
[1]Student, [2]Student, [3]Student, [4]Student
[1]Department Of Computer Engineering,
[1]All India Shri Shivaji Memorial Society's Institute Of Information Technology, Pune, India

Abstract— The ultimate goal of individuals seeking the help from web is to acquire a particular set of data regarding a domain through collaborative environment. In an organisation, the employees or higher authorities may require to work with some business insight software or purchase the same, for this many must have referred the elements or products online. The fine grained knowledge acquired through their surfing
may be shared with the employees to know about the software and share the learned knowledge. We perform the dissection of individuals web surfing history to get the fine grained knowledge. Following are the two stages in which fine grained learning is mined.
1. A non parametric generative model is used to prepare sets of web surfing data.
2. A original discriminative Support Vector Machine(SVM) is created to mine fine grain information in every project. To find proper individuals, those who are sharing information, the excellent master enquiry technique is connected to get mined results. To establish the fine grained overview of mining system the probes web
surfing information is gathered from the browser. When it is enhanced or joined with master hunt, the precision grows notably in relation with applying the fantastic master technique straight forwardly on web surfing data.

Index Terms— Advisor Search ,Fine Grained Knowledge Sharing, Text Mining Graphical Modules, Non Parametric Generative Models, Collaborative Environment ,Client-Server Model.

I. INTRODUCTION

PROJECT IDEA

With the web and with partners/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 comparative 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) programming, and representatives from these divisions may have concentrated on online about diverse BI instruments and their elements freely. In an examination lab, individuals are regularly centered around tasks which require comparable foundation information. 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, contrasted with the web.

For the first situation, it is more profitable for a worker to get advices on the decisions of BI devices and clarifications of their components from experienced representatives; for the second situation, the first analyst could get proposals on model configuration and great taking in materials from the second scientist .A great many people in synergistic situations would be glad to impart encounters to and offer recommendations to others on particular issues. On the other hand, discovering a perfect individual is testing because of the assortment of data needs. In this paper, we explore how to empower such learning sharing system by dissecting client information.

PROBLEM STATEMENT

We examine fine-grained knowledge sharing in community oriented situations. We propose to dissect individuals web surfing Information to compress the fine grain learning gained by them.

MOTIVATION OF THE PROJECT

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, contrasted with the web.

II. GOALS AND OBJECTIVES

1. To find people who are most likely having the desired knowledge of fine grained knowledge.
2. To deal with advisor search by exploiting the data generated from users past online surfing.
3. To sum up fine-grained aspects which can provide a fine grained description of the knowledge gained by a person.
4. To provide a vigorous sharing environment.
5. To provide easy access of desired information and save time of repetitive efforts.

III. LITERATURE SURVEY

The Infinite Hidden Markov Model
Author: Matthew J. Beal Zoubin Ghahramani Carl Edward Rasmussen
We demonstrate that it is conceivable to extend hidden Markov models to have a countably endless number of hidden states. By utilizing the hypothesis of Dirichlet forms we can verifiably incorporate out the boundlessly numerous move parameters, leaving just three hyper parameters which can be gained from data. These three hyper parameters character a various leveled .Dirichlet process equipped for catching a rich arrangement of transition dynamics.The three hyper parameters control the time size of the motion, the sparsity of the fundamental state-move framework.

Formal Models for Expert Finding in Enterprise Corpora
Author: Krisztian Balog, Leif Azzopardi
Searching an association's report vaults down specialists gives a cost effective solution for the task of expert finding. We show two general methodologies to master seeking given a report accumulation which are formalized utilizing generative probabilistic models. The main of these straightforwardly models a specialist's learning taking into account the archives that they are connected with, whilst the second finds reports on theme, and after that discovers the related master. Framing dependable affiliations is pivotal to the execution of master discovering frameworks.

Hierarchical Topic Models and the Nested Chinese Restaurant Process
Author: David M. Blei Thomas L. Griffiths
We address the issue of taking in point chains of command from data. We take a Bayesian methodology, producing a proper earlier through a conveyance on parcels that we allude to as the settled Chinese restaurant process. This nonparametric former permits discretionarily substantial fanning components and promptly suits growing data collections. We assemble a progressive theme model by consolidating this earlier with a probability that depends on a progressive variation of inactive Dirichlet distribution.

Variational Inference for Dirichlet Process Mixtures
Author: David M. Blei
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. However, MCMC sampling can be prohibitively slow, and it is important to explore alternatives. One class of alternatives is provided by variational methods, a class of deterministic algorithms that convert inference problems into optimization problems. In this paper, we present a variational inference algorithm for DP mixtures.

Latent Dirichlet Allocation
Author: David M. Blei, Andrew Y. Ng
We depict inactive Dirichlet allotment (LDA), a generative probabilistic model for accumulations of discrete data, for example, content corpora. LDA is a three-level progressive Bayesian model, in which each thing of a gathering is displayed as a limited blend over a hidden arrangement of points. Every subject is, in turn, displayed as a vast blend over a basic arrangement of subject probabilities. In the setting of content displaying, the theme probabilities give
an unequivocal representation of a record.

Latent Dirichlet Allocation
Author: David M. Blei
We describe latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a
finite mixture over an underlying set of topics. Each topic is, in turn, modelled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation
of a document. We present efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation.

Dynamic Topic Models
Author: John D. Lafferty
A family of probabilistic time series models is developed to analyze the time evolution of topics in large document collections. The approach is to use state space models on the natural parameters of the multinomial distributions that
represent the topics. Variational approximations based on Kalman filters and nonparametric wavelet regression are developed to carry out approximate posterior inference over the latent topics. In addition to giving quantitative, predictive models of a sequential corpus, dynamic topic models provide a qualitative
window into the contents of a large document collection.

Formal Models for Expert Finding on DBLP Bibliography Data
Author: Irwin King
Finding relevant experts in a specific field is often crucial for consulting, both in industry and in academics. The aim of this paper is to address the expert finding task in a real world academic field. We present three models for expert
finding based on the large-scale DBLP bibliography and Google Scholar for data supplementation. The first, a novel weighted language model, models an expert candidate based on the relevance and importance of associated documents
by introducing a document prior probability, and achieves much better results than the basic language model. The second, a topic-based model, represents each candidate as a weighted sum of multiple topics, whilst the third,
a hybrid model, combines the language model and the topic-based model.

Overview of the TREC Enterprise Track
Author: Arjen P. de
The goal of the enterprise track is to conduct experiments with enterprise data intranet pages, email archives, document repositories that reflect the experiences of users in real organizations, such that for example, an email ranking
technique that is effective here would be a good choice for deployment in a real multi-user email search application. The enterprise track began this year as the successor to the web track, and this is reflected in the tasks and measures.
While the track takes much of its inspiration from the web track, the foci are on search at the enterprise scale, incorporating non-web data and discovering relationships between entities in the organisation.

Working Knowledge: How Organizations Manage What They Know
Author: Thomas H. Davenport
We start with those more familiar terms both because they are more familiar and because we can understand knowledge best with reference to them. Confusion about what data, information, and knowledge are how they differ,
what those words mean – has resulted in enormous expenditures on technology initiatives that rarely deliver what the firms spending the money needed or thought they were getting. Often firms don't understand what they need
until they invest heavily in a system that fails to provide it. However basic it may sound, then, it is still important to emphasize that data, information, and knowledge are not interchangeable concepts.

IV. EXISTING SYSTEM

Expert search aims at retrieving people who have expertise on the given query topic. 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 Balog et al. proposed a language model framework for expert search. 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. This process was formulated in a generative probabilistic model. Balog et al.showed that Model 2 performed better and it became one of the most prominent methods for expert search. Other methods have been proposed for enterprise expert search but the nature of these methods is still accumulating relevance scores of associated documents tocandidates. Expert retrieval in other scenarios has also been studied, e.g. online question answering communities ,academic society .

Disadvantages of this method:

1) An analyst might need to tackle an information mining issue utilizing nonparametric graphical models which she is not acquainted with but rather have been concentrated on by another analyst some time recently.
2) Here we may have to go through repetitive data which may consume our time and may even compromise with the quality of Information.

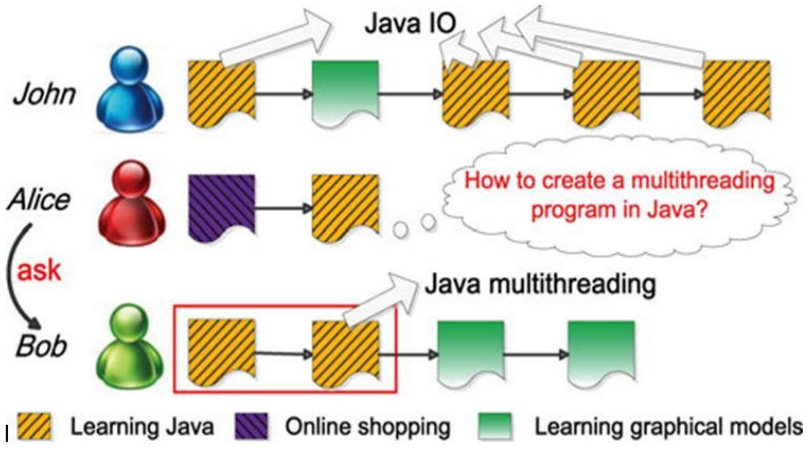
V. PROPOSED SYSTEM

PROPOSED WORK

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 different from documents in enterprise repositories. A person typically generates multiple sessions for a microaspect 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 (same task but other microaspects) 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 microaspects and show the superiority of our search scheme over the simple idea of applying traditional expert search methods on session data directly.

Advantages of the proposed system:

1) Web surfing information is grouped into assignments by a nonparametric generative model.
2) A novel discriminative limitless Hidden Markov Model is created to mine fine-grained angles in every undertaking.



LATENT DIRICHLET ALLOCATION ALGORITHM

(LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. LDA is an example of a topic model and was first presented as a graphical model. 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 .

CLUSTERING BY K-MEANS ALGORITHM

k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster.

Algorithm steps are:

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

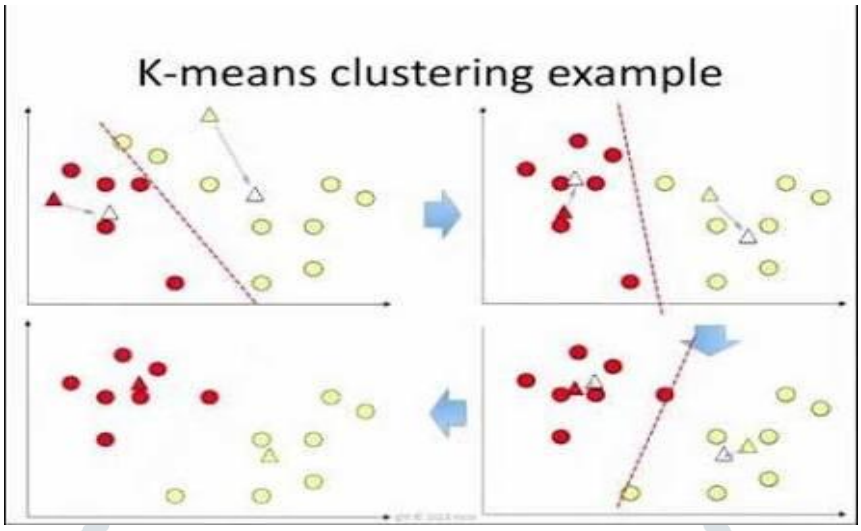
1) Randomly select 'c' cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, 'ci' represents the number of data points in ith cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

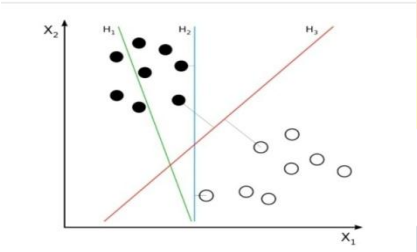
6) If no data point was reassigned then stop, otherwise repeat from step 3).



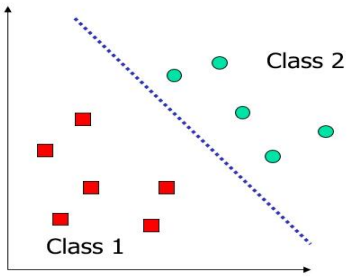
TEXT CLASSIFICATION USING SUPPORT VECTOR MACHINE

Support vector machines(SVMs, also support vector networks are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. It is non-probabilistic binary linear classifier. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other.

Support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. A good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin).

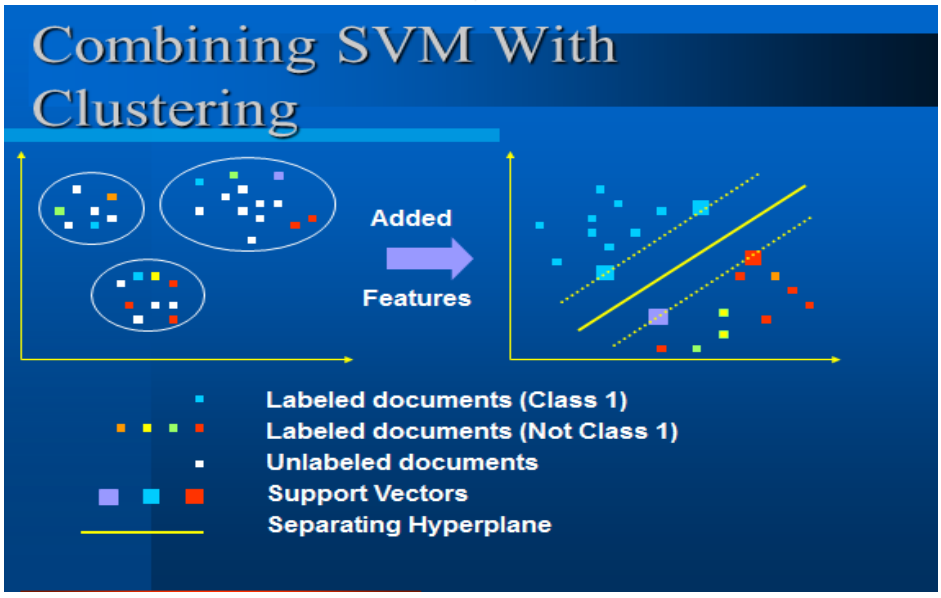


H_1 does not separate the classes. H_2 does, but only with a small margin. H_3 separates them with the maximum margin.



MINING FINE GRAIN KNOWLEDGE

Mining of fined grained knowledge is done by combining the process of clustering and then classifying the clustered data using SVM.



PAGE RANKING ALGORITHM

Since the early stages of the world wide web, search engines have developed different methods to rank web pages. For the purpose of better search results and especially to make search engines resistant against automatically generated web pages based upon the analysis of content specific ranking criteria (doorway pages), the concept of link popularity was developed. Following this concept, the number of inbound links for a document measures its general importance. Hence, a web page is generally more important, if many other web pages link to it. Contrary to the concept of link popularity, PageRank is not simply based upon the total number of inbound links. The basic approach of PageRank is that a document is in fact considered the more important the more other documents link to it, but those inbound links do not count equally.A document ranks high in terms of PageRank, if other high ranking documents link to it. So, within the PageRank concept, the rank of a document is given by the rank of those documents which link to it.

HIERARCHICAL CLUSTERING ALGORITHM

Hierarchical clustering algorithm is of two types:
i) Agglomerative Hierarchical clustering algorithm or AGNES (agglomerative nesting) and
ii) Divisive Hierarchical clustering algorithm or DIANA (divisive analysis).
Both this algorithm are exactly reverse of each other. So we will be covering Agglomerative Hierarchical clustering algorithm in detail.
Agglomerative Hierarchical clustering -This algorithm works by grouping the data one by one on the basis of the nearest distance measure of all the pairwise distance between the data point. Again distance between the data point is recalculated but which distance to consider when the groups has been formed? For this there are many available methods. Some of them are:
1) single-nearest distance or single linkage.
2) complete-farthest distance or complete linkage.
3) average-average distance or average linkage.
4) centroid distance.
5) ward's method - sum of squared euclidean distance is minimized.
This way we go on grouping the data until one cluster is formed.

VI. CONCLUSION AND FUTURE SCOPE

We presented a novel issue, fine-grained knowledge sharing in cooperative situations, which is alluring in rehearse. We recognized uncovering finegrained knowledge reflected by individuals' associations with the outside world as the way to tackling this issue. We proposed a two-stage system to mine fine-grained knowledge and coordinated it with the fantastic master search system for discovering right guides. Probes genuine web surfing data appeared empowering results. There are open issues for this issue. The fine grained knowledge could have a various leveled structure. For sample, "Java IO" can contain "Document IO" and "System IO" as sub-knowledge. We could iteratively apply d-iHMM on the scholarly small scale angles to determine a chain of command, yet how to look over this pecking order is not an inconsequential issue.

REFERENCES

[1] 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, 2012, pp. 4350.

[2] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen, The infinite hidden Markov model, in Proc. Adv. Neural Inf. Process. Syst.,2011, pp. 577584.

[3] M. Belkin and P. Niyogi, Laplacian Eigenmaps and spectral techniques for embedding and clustering, in Proc. Adv. Neural Inf. Process. Syst., 2001, pp. 585591.

[4] D. Blei and M. Jordan, Variational inference for Dirichlet process mixtures, Bayesian Anal., vol. 1, no. 1, pp. 121143, 2014.

[5] 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., 2013, pp. 1724.

[6] D. M. Blei and J. D. Lafferty, Dynamic topic models, in Proc. Int. Conf. Mach. Learn., 2007, pp. 113120.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res., vol. 3, pp. 9931022, 2006.

[8] P. R. Carlile, Working knowledge: How organizations manage what they know, Human Resource Planning, vol. 21, no. 4, pp. 5860, 2010.

[9] N. Craswell, A. P. de Vries, and I. Soboroff, Overview of the TREC 2005 enterprise track, in Proc. 14th Text REtrieval Conf.,2009, pp. 199205.

[10] H. Deng, I. King, and M. R. Lyu, Formal models for expert finding on DBLP bibliography data, in Proc. IEEE 8th Int. Conf. Data Mining, 2008, pp. 163172.