IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

MICRO ASPECT MINING IN A COLLABORATIVE ENVIRONMENT USING A NOVEL DISCRIMINATIVE INFINITE HIDDEN MARKOV MODEL

M.Swapnaa*, Mr. M. Kalidass

[*] PG Scholar, Department of Computer Science and Engineering, R.V.S. Educational Trust's Group of Institutions, Dindigul, Tamil Nadu

Assistant Professor, Department of Computer Science and Engineering, R.V.S. Educational Trust's Group of Institutions, Dindigul, Tamil Nadu

DOI: 10.5281/zenodo.51455

ABSTRACT

Collaborative environments, which enable companywide global teams to identify the source of the problem and develop a response, are an excellent antidote to a lack of preparedness. Knowledge sharing is an activity through which knowledge is exchanged among people, friends, families, communities or organizations. In order to gain knowledge, user may try to acquire similar information on the web in this collaborative environment. The framework formulates tasks from sessions. There is no existing technique for micro aspect mining. A novel discriminative infinite Hidden Markov Model is proposed to mine micro aspects and evolution patterns in each task. The goal is not finding domain experts but a person who has the desired specific knowledge. In this project first summarizing web surfing data into fine grained aspects, and then search over these aspects. This strategy is compared with searching advisors directly over sessions both analytically and empirically.

KEYWORDS: Advisor search, text mining, Dirichlet processes, graphical models.

INTRODUCTION

Interacting with the web and with partners, friends to acquire information is a daily routine of many human beings. Web search engines attempt to satisfy the users information needs by ranking web pages with respect to queries. But the actuality of web search is that it is a process of querying, learning, and reformulating. A series of interactions between user and search engine is necessary to satisfy a single information need. *In times of crisis or catastrophe, managers have regretted for not having more efficient, responsive communication systems.*

Collaborative environments, which enable companywide global teams to recognize the source of the problem and develop a response, are an excellent antidote to a need of preparedness. These authors narrate how collaborative environments can do nothing less than save an organization from disaster.

In this era of global connectivity, organizations are increasingly adopting and applying CEs to tap into the knowledge and skill of their employees, customers and business partners. They are constructed from a range of computer and communications technologies, such as messaging, e-mail, chat rooms, discussion databases, mobile communicators, media spaces, shared whiteboards, cyber cafes, and audio, video or web conferences.

Perhaps of greater significance, the use of collaborative environments is increasing in response to reduced travel budgets, international terrorism, and world health epidemics such as SARS, and wild events such as the Great Blackout of 2003 that paralyzed over 50 million people in the eastern United States and Canada.

Organizations have recognized that knowledge constitutes a valuable invisible asset for creating and encourage competitive advantages. Knowledge sharing activities are generally supported by knowledge management systems. However, technology comprises only one of the many factors that affect the sharing of knowledge in organizations, such as organizational culture, trust, and motivation.

The sharing of knowledge constitutes a major challenge in the field of knowledge management because some employees tend to withstand sharing their knowledge with the rest of the organization. The system proposes a two step framework for mining fine grained knowledge. In the first step, it formulate tasks from sessions and design an infinite Gaussian mixture model based on DP [9] to cluster sessions and also adopt a nonparametric scheme since the number of tasks is difficult to predict.

Knowledge and data engineering stimulates the exchange of ideas and interaction between these two related fields of interest. *KDE* reaches a worldwide audience of researchers, designers, managers and users. The major aim is to identify, investigate and examine the underlying principles in the design and effective use of these systems. *KDE* achieves this aim by publishing original research results, technical approach and news items concerning data engineering, knowledge engineering, and the interface of these two fields.

While the expert may consciously persuasive some parts of his or her knowledge, he or she will not be aware of a significant part of this knowledge since it is hidden in his or her skills. This knowledge is not directly available, but has to be built up and structured during the knowledge acquisition phase. Therefore, this knowledge acquisition process is no longer seen as a transfer of knowledge into an appropriate computer representation, but as a model construction process. The modeling process is a cyclic process. New observations may lead to a refinement, modification or execution of the already built up model.

The rest of the paper is organized as follows: the next section outlines related work. Section 3 gives the proposed system present the Gaussian DP model for clustering sessions into tasks also describes the proposed d-iHMM model for mining fine-grained aspects in each task (i.e. session cluster), and finally, Section 4 concludes our work.

RELATED WORK

Michael C. Hughes and Erik B. Sudderth have proposed Memoized Online Variational Inference for Dirichlet Process Mixture Models [6] in the year 2013, that the variational inference algorithms provide the most effective framework for large scale training of Bayesian nonparametric models. Stochastic online approaches are promising, but are sensitive to the chosen learning rate and often converge to poor local optima.

They present a new algorithm, memoized online variational inference, which scales to very large datasets while avoiding the complexities of stochastic gradient. The algorithm maintains finite dimensional sufficient statistics from batches of the full dataset, requiring some additional memory but still scaling to millions of examples.

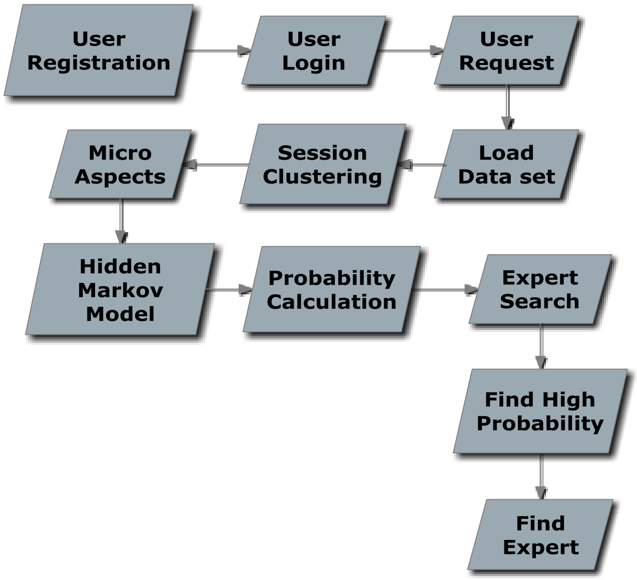
Anil K. Jain has proposed Data clustering 50 years beyond K-means [1] in the year 2010, organizing data into sensible groupings is one of the most fundamental modes of understanding and learning. A common scheme of scientific classification puts organisms into a system of ranked taxa domain, kingdom, phylum, class, etc. Cluster analysis is the formal study of methods and algorithms for grouping, or clustering, objects according to measured or perceived intrinsic characteristics or similarity.

Ognjen Arandjelovic has proposed Discriminative k-Means Clustering [7] in the year 2010, K-means algorithm is a partitioned clustering method. Over 60 years old, it has been successfully used for a variety of problems. The popularity of k-means is in large part a consequence of its simplicity and efficiency.

Yi Fang and Luo Si have proposed Discriminative Models of Integrating Document Evidence and Document Candidate Associations for Expert Search [10] in the year 2010, the key ingredient in these methods is to determine associations between people and documents because the associations are ambiguous in the TREC scenarios as well as in many realistic settings. Previous works have investigated different metrics or a combination of them to measure the associations, but the way of choosing or combining them is rather often heuristic and lacks of a clear justification.

Hongbo Deng and Irwin King et al have proposed Formal Models for Expert Finding on DBLP Bibliography Data [2] in the year 2009, finding relevant experts in a specific field is often crucial for consulting, both in industry and in academia. The aim of this project is to address the expert finding task in a real world academic field, and to present three models for expert finding based on the large scale DBLP bibliography and google scholar for data supplementation.

PROPOSED SYSTEM



System Architecture

In proposed system, a fine grained knowledge sharing is proposed in collaborative environments. Web surfing data to analyze the members and to summarize the fine grained knowledge acquired by them is proposed in the system. The framework has two steps. First, web surfing data is clustered into tasks by a nonparametric generative model second, a novel discriminative infinite Hidden Markov Model is developed to mine fine grained aspects in each task and to find proper members for knowledge sharing, the classic expert search method is applied to the mined results.

A two step framework for mining fine grained knowledge is proposed by the system. In the first step, the system formulates tasks from sessions. The system design an infinite Gaussian mixture model based on DP [9] to cluster sessions. The system adopts a nonparametric scheme since the number of tasks is difficult to predict. The system then extracts micro aspects from sessions in each task.

As shown in fig 3.1 the challenges are the number of micro aspects in a task is unknown; sessions for different micro aspects of a task are textually similar; sessions for a micro aspect might not be successive. To this end, a novel discriminative infinite Hidden Markov Model is proposed to mine micro aspects and evolution patterns in each task. A background model is introduced in order to strengthen the discriminative power. Finally, the systems apply a language model based expert search method over the mined micro aspects for advisor search.

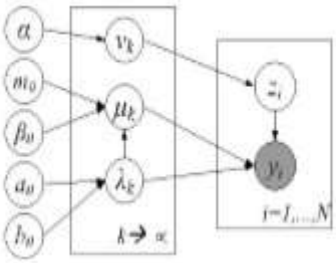
Session Clustering

The input of this step is W, where each w_i is a D_0=1 word frequency vector with D_0 as the vocabulary size. The intuition is that contents generated for the same task are textually similar while those for different tasks are dissimilar. Hence, clustering is a natural choice for recovering tasks from sessions. In our case, it is difficult to preset the number of tasks given a collection of sessions. Therefore, we need to automatically determine the number of clusters (k), which is also one of the most difficult problems in clustering research. Most methods for automatically determining k run the clustering algorithm with different values of k and choose the best one according to a predefined criterion [1], which could be costly.

In this work, we advocate using a generative model with a Dirichlet Process prior [9] for clustering. DPs provide nonparametric priors for *k* and the most likely *k* is learned automatically. A DP, written as $G\sim DP(\alpha, G_0)$, can be interpreted as drawing components (clusters here) from an infinite component pool, with α called the scaling parameter and G_0 being the prior for a random component. An intuitive interpretation of DP is the stick-breaking construction:

$$\pi_i(v)=v_i\prod_{j=1}^{i-1}(1-v_j),\ G=\sum_{i=1}^{\infty}\pi_i\delta_{\psi_i}$$

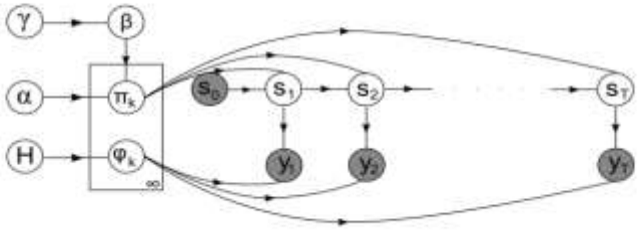
where $v=\{v_1,v_2,...\}$ with each v_i drawn from the Beta distribution Beta $(1,\alpha)$, ψ_i is a component drawn from G_0. π_i is the mixture weight of ψ_i given by breaking the current length of the "stick" by the fraction v_i. The generation of π is often written asπ ~ GEM(α). π defines a prior mixing distribution among the infinite many components. The posterior mixing distribution and the real number of components drawn from the DP is then learned from the data.



Gaussian Dirichlet Process mixture model.

The graphical representation of GDP is depicted in figure. The DP prior is represented by the stick-breaking construction process. When using probabilistic models for clustering, the Gaussian mixture model is a common choice and can be viewed as a probabilistic version of k-means [1]. However, the data dimensionality D_0 is too high to apply Gaussian distributions in our case. Therefore, we first apply the well known Laplacian Eigenmap (LE) technique [4] to reduce the dimensionality from D_0 to D where D_0>>D. We choose LE since it could also capture the nonlinear manifold structure of a task, e.g. the topics of a task could evolve and drift which could be described by the half-moon structure [4].

Mining Fine-Grained Knowledge



Graphical representation of iHMM

The major challenge of mining micro-aspects is that the micro-aspects in a task are already similar with one another. If we model each component separately, it is likely that we mess up sessions from different micro-aspects, i.e. leading to bad discrimination. Therefore, we should model different micro-aspects in a task jointly, separating the common content characteristics of the task from the distinctive characteristics of each micro-aspect. To this end, we extends the infinite Hidden Markov Model (iHMM) [5] and propose a novel discriminative infinite Hidden Markov Model to mine micro-aspects and possible evolution patterns in a task. The graphical representation is shown in figure d-iHMM is fundamentally based on the Hierarchical DP model [11] where the infinite component pools (corresponding to $\{\pi_k\}$)

of all states share the same base infinite component pool (corresponding to β). The generative process is summarized as follows:

1) Draw $\beta|\gamma_0 \sim \mathrm{GEM}(\gamma_0)$
2) For $k = 1, 2, \ldots$, draw $\pi_k \sim \mathrm{DP}(\alpha, \beta)$, $\theta_k \sim \mathrm{Dir}(h_0, \ldots, h_0)$
3) For $t = 1$ to T:
   a) Draw $s_t|s_{t-1} \sim \mathrm{Mult}(\pi_{s_{t-1}})$, $\eta_t|\delta_0 \sim \mathrm{Beta}(1, \delta_0)$
   b) For each word w_{ti} in the tth session
      i) Draw $z_{ti}|\eta_t \sim \mathrm{Bernoulli}(\eta_t)$
      ii) If $z_{ti} = 1$, draw $w_{ti} \sim \mathrm{Mult}(\theta_{s_t})$, else draw $w_{ti} \sim \mathrm{Mult}(\theta_b)$.

The beam sampling method for iHMM is proposed in [3], which is shown to converge to the true posterior much faster than a classical Gibbs sampler. Therefore, we develop a beam sampler for our d-iHMM model. Beam sampling adopts the slice sampling [8] idea to limit the number of states considered at each time step to a finite number, so that dynamic programming can be used to sample whole state trajectories efficiently.

Each sampling iteration samples $\{u_t\}$, $\{s_t\}$, $\{z_{ti}\}$, $\{\theta_k\}$, $\{\pi_k\}$ and β in turn. We first sample $\{u_t\}$ as described above and create more states if the maximum unassigned probabilities in $\{\pi_k\}$ (the last element of each π_k) is greater than the minimum of $\{u_t\}$. Then we perform a forward sweep of $\{s_t\}$ where for the tth step we compute

$$\begin{aligned} & p(s_t|w_{1:t}, u_{1:t}, z_t) \\ & \quad \propto p(s_t, u_t, w_t|w_{1:t-1}, u_{1:t-1}, z_t) \\ & \quad = p(w_t|s_t, z_t) \\ & \qquad \times \sum_{s_{t-1}} \mathbb{I}(u_t < \pi_{s_{t-1}s_t}) p(s_{t-1}|w_{1:t-1}, u_{1:t-1}, z_{t-1}), \end{aligned} \tag{1}$$

Then we perform a backward sweep to sample each s_t given s_{t+1} by

$$\begin{aligned} & p(s_t|s_{t+1}, w_{1:T}, u_{1:T}, z_t) = p(s_t|s_{t+1}, w_{1:t}, u_{1:t+1}, z_t) \\ & \propto p(s_t|w_{1:t}, u_{1:t+1}, z_t) p(s_{t+1}|s_t, u_{t+1}) \\ & \propto p(u_{t+1}|s_t) p(s_t|w_{1:t}, u_{1:t}, z_t) \frac{p(u_{t+1}|s_t, s_{t+1}) p(s_{t+1}|s_t)}{p(u_{t+1}|s_t)} \\ & = p(s_t|w_{1:t}, u_{1:t}, z_t) \mathbb{I}(u_{t+1} < \pi_{s_t s_{t+1}}). \end{aligned} \tag{2}$$

In order to efficiently sample z_{ti}, we integrate out $\{\eta_t\}$.This makes all z_{ti}'s belonging to w_t dependent on one another. Let z_{ti} be the set of z variables for w_t except z_{ti}. We have

$$\begin{aligned} & p(z_{ti}|z_{-ti}, \delta_0, w_t, \theta, \theta_b, s_t) \\ & \quad \propto p(z_{ti}, z_{-ti}, w_t|\delta_0, \theta, \theta_b, s_t) = p(z_t|\delta_0) p(w_t|z_t, s_t, \theta, \theta_b) \\ & \quad \propto \frac{\mathrm{B}(\sum_j z_{tj} + 1, |w_t| - \sum_j z_{tj} + \delta_0)}{\mathrm{B}(1, \delta_0)} p(w_{ti}|z_{ti}, s_t, \theta, \theta_b), \end{aligned} \tag{3}$$

where $|w_t|$ is the number of words in w_t. The final sampling probability ratio is

$$\frac{p(z_{ti} = 1|\cdots)}{p(z_{ti} = 0|\cdots)} = \frac{p(w_{ti}|\theta_{s_t})(\sum_{j \neq i} z_{tj} + 1)}{p(w_{ti}|\theta_b)(|w_t| - \sum_{j \neq i} z_{tj} + \delta_0 - 1)}, \tag{4}$$

where $p(w|\theta_k)$ means the probability of generating word w by θ_k.

CONCLUSION AND FUTURE WORK

The system has proposed a fine grained knowledge sharing in collaborative environments to sharing knowledge. The system has a two step framework to mine fine grained knowledge and integrated it with the classic expert search method for finding right advisors. The system formulated the tasks from sessions and then designed an infinite Gaussian mixture model based on Dirichlet process to cluster sessions. Then extracted micro aspects from sessions in each task. Finally, a novel discriminative infinite Hidden Markov Model has proposed to mine micro aspects and evolution in each task. It is important to recognize the semantic structures and summarize the session data into micro aspects so that find the desired advisor accurately. Experiments on both datasets show that the scheme is effective and outperforms the one using raw session data.

There are some issues which may be rectified in the future work. They are the fine grained knowledge could have a hierarchical structure. The method is applied iteratively on the learned micro aspects to derive a hierarchy, but how to search over this hierarchy is not a trivial problem. The basic search model can be refined that is the accuracy is improved in expert search phase using Baum Welch algorithm. Privacy is also an issue therefore security is may be implemented in future work.

REFERENCES

[1] K. Jain, Data clustering: 50 years beyond k-means, Pattern Recog. Lett., vol. 31, no. 8, pp. 651-666, 2010.
[2] 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, 2009, pp. 163-172.
[3] J. Van Gael, Y. Saatci, Y. Teh, and Z. Ghahramani, Beam sampling for the infinite hidden Markov model, in Proc. Int. Conf. Mach. Learn., 2008, pp. 1088-1095.
[4] 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.
[5] 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.
[6] Michael C. Hughes and Erik B. Sudderth, Memoized Online Variational Inference for Dirichlet Process Mixture Models, in Proc. Adv. Neural Inf. Process. Syst., 2013, pp. 17-24.
[7] Ognjen Arandjelovic, Discriminative k-Means Clustering, Int. Conf. on neural net., vol.1, 2010.
[8] R. M. Neal, Slice sampling, Ann. Statist., vol. 31, pp. 705-741, 2003.
[9] T. S. Ferguson, A Bayesian analysis of some nonparametric problems, Ann. Statist., vol. 1, no. 2, pp. 209-230, 1973.
[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] Y. Teh, M. Jordan, M. Beal, and D. Blei, Hierarchical Dirichlet processes, J. Am. Statist. Assoc., vol. 101, no. 476, pp. 1566-1581, 2006.