ISSN 2395-1621

Text Mining using Hybrid Algorithms

[#1]Trushagni Bhoi, [#2]Sayali Shinde, [#3]Kajal Kajale, [#4]Prof. Nitin Shivale

[1]trushagni65@gmail.com
[2]sayalishinde187@gmail.com
[3]kkajale10@gmail.com

[#123]Department of Computer Engineering
JSPM's
Bhivarabai Sawant Institute Of Technology Research
Wagholi, Pune.

ABSTRACT

In collaborative environments, members may try to acquire similar information on the web in order to gain knowledge in one domain. For example, in a company several departments may successively need to buy business intelligence software and employees from these departments may have studied online about different business intelligence tools and their features independently. It will be productive to get them connected and share learned knowledge. We investigate fine-grained knowledge sharing in collaborative environments. We propose to analyze members' web surfing data to summarize the fine-grained knowledge acquired by them. A two-step framework is proposed for mining fine-grained knowledge: (1) web surfing data is clustered into tasks by a nonparametric generative model. (2) a novel discriminative infinite Hidden Markov Model is developed to mine fine-grained aspects in each task. Finally, the classic expert search method is applied to the mined results to find proper members for knowledge sharing. Experiments on web surfing data collected from our lab at UCSB and IBM show that the fine-grained aspect mining framework works as expected and outperforms baselines. When it is integrated with expert search, the search accuracy improves significantly, in comparison with applying the classic expert search method directly on web surfing data.

Keywords: **Advisor search, Text mining, Dirichlet processes, Graphical models.**

ARTICLE INFO

Article History

Received: 18th October 2015

Received in revised form :20th October 2015

Accepted : 22nd October 2015

Published online : 24th October 2015

I. INTRODUCTION

While communicating with the web and with colleagues/friends to acquire information is a daily routine of many human beings. In a collaborative environment, it could be common that members try to acquire similar information on the web in order to gain specific knowledge in one domain. For example in a company several departments may successively need to buy business intelligence (BI) software and employees from these departments may have studied online about different BI tools and their features independently.

In these cases, resorting to a right person could be far more efficient than studying by oneself, since people can provide digested information, insights and live interactions, com- pared to the web. For the first scenario, it is more productive for an employee to get advices on the choices of BI tools and explanations of their features from experienced employees; for the second scenario, the first researcher could get suggestions on model design and good learning materials from the second researcher. 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.

The scene is, Alice starts to surf the web and wants to learn how to develop a Java multithreading program, which has already been studied by Bob (red rectangle). In this case, it might be a good idea to consult Bob, rather than studying by herself. We aim to provide such recommendations by analyzing surfing activities automatically. In this example, not necessarily Bob is an expert in every aspect of Java programming; how- ever, due to his significant surfing activities in Java multi- threading, it is reasonable to assume that he has gained enough knowledge in this area so that he can help Alice . Even if Bob is still learn- ing, he could share his experiences in learning and possibly suggest good

learning materials to Alice, thus saving Alice's effort and time. This scenario departs from the traditional expert search problem in that expert search aims to find domain experts based on their associated documents in an enterprise repository, while 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 example, in the above example John spent a lot of effort on "Java IO" which is only partially relevant to Alice's need. If we simply treat web surfing data as a collection of documents and apply traditional expert search methods, John would be ranked higher than Bob since he viewed more contents about "Java", though not quite relevant. Therefore, it is necessary to first summarize people's fine-grained knowledge reflected in their web surfing activities by recognizing the semantic structures, and then search over the mined pieces of fine-grained knowledge (e.g. "Java IO"). We call this search scenario, "advisor search", to differentiate from traditional expert search. We use the term advisor search to emphasize that the knowledge of the retrieved people might not be very deep, but good enough to help others if they have not solidly gained the related knowledge yet.

We define a session as an aggregation of consecutively browsed web contents of a user that belong to the same task. Sessions are atomic units in our analysis. The content of sessions in a task could evolve gradually: 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. 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 is critical: it can provide a detailed description of the knowledge gained by a person, which is the basis for advisor search.

We propose a two-step framework for mining fine-grained knowledge (micro-aspects):-

(1) In the first step, we formulate tasks from sessions. We design an infinite Gaussian mixture model based on Dirichlet Process (DP) [1] to cluster sessions. We adopt a nonparametric scheme since the number of tasks is difficult to predict.

(2) We then extract micro-aspects from sessions in each task. 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 consecutive. To this end, a novel discriminative infinite Hidden Markov Model (d-iHMM) is proposed to mine micro-aspects and evolution patterns (if any) in each task. A background model is introduced in order to enhance the discriminative power. Finally, we apply a language model based expert search method [1] over the mined micro-aspects for advisor search.

II. LITERATURE SURVEY

T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," We demonstrate the use of a variant of the nonparametric Bayesian (NPB) forward-backward (FB) method for sampling state sequences of hidden Markov models (HMMs), when the continuous-valued observations follow autoregressive (AR) processes.. Instead one uses hierarchical Dirichlet processes (HDPs) as priors for the state-transition probabilities to account for a potentially infinite number of states. We show that by approximately integrating out some parameters of the model, one can alleviate this problem considerably.

Drawbacks-A Bayesian analysis approach is fruitful in many ways but it has rather been unsuccessful in terms of non-parametric problems.

H. Deng, I. King, and M. R. Lyu, "Formal models for expert finding on DBLP bibliography data". Finding relevant experts in a specific field is often crucial for consulting, both in industry and in academia. The aim of this paper was to address the expert-finding task in a real world academic field. Three models for expert finding based on the large-scale DBLP bibliography and Google scholar for data supplementation are presented. We evaluate our system using a benchmark dataset based on human relevance judgments of how well the expertise of proposed experts matches a query topic. Evaluation results show that the hybrid model outperforms other models in few metrics.

Drawbacks:-Evaluation results show that the hybrid model outperforms other models only in particular types of metrics but failed to work in all.

III. PROPOSED SYSTEM

We In this section we review research fields that are related to our work:. expert search, analysis of user search tasks and topic modeling:

1. Expert search

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.

In this paper we develop nonparametric generative models to mine micro aspects and show the superiority of our search scheme over the simple idea of applying traditional expert search methods on session data directly.

2. Analysis of Search Tasks

Recently researchers have focused on detecting, modelling and analysing user search tasks from query logs. The search tasks are interleaved and used classifiers to segment the sequence of user queries into tasks.

First, we consider general web surfing contents (including search), rather than search engine query logs. Query logs do not record the subsequent surfing activity after the user clicked a relevant search result. Moreover, it is found that 50 percent of a user's online pageviews are content browsing. Web surfing data provides more comprehensive

information about the knowledge gaining activities of users. Although various methods were proposed for extracting search tasks in query logs, these methods cannot be applied in our setting since they exploit query log specific properties. Second, none of the above works tried to mine fine-grained aspects for each task. When studying, people could spend some effort on one fine-grained aspect of a task and generate multiple contents. Summarizing fine-grained aspects can provide a fine- grained description of the knowledge gained by a person. Finally, none of existing works which analyse user online behaviours address advisor search by exploiting the data generated from users' past online behaviours.

3. Topic Modelling

Topic modelling is a popular tool for analysing topics in a document collection. The most prevalent topic modelling method is Latent Dirichlet Allocation (LDA) .Based on LDA, various topic modelling 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 . However, our problem is not a topic modelling problem.

Our goal is to recover the semantic structures of people's online learning activities from their web surfing data, i.e. identifying groups of sessions representing tasks (e.g. learning "Java") and micro-aspects (e.g. learning "Java multithreading"). While topic modelling decomposes a document into topics. After applying topic modelling methods on session data, it is still difficult to find the right advisor by using the mined topics. This is because a person with many sessions containing par- tially relevant topics would still be ranked unexpectedly high, due to the accumulation of relevance among sessions. Grouping sessions into micro-aspects is important for advisor search.

IV. PROPOSED ALGORITHM

K-Means Algorithm

- K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem.
- K-means algorithm is used for clustering the data into the groups/objects based on similar properties.

Advantages:

1) Fast and robust
2) Easier to understand
3) Relatively efficient
4) Gives best result when data set are distinct or well separated from each other.

Ranking Algorithm

- The original Page Rank algorithm was described by Lawrence Page and Sergey Brin .
- Page rank was named after Larry page, one of the founder of Google.
- Ranking algorithm is used for rank websites in their search engine result.
- Page rank is the way of measuring the importance of website pages.

V. CONCLUSION

We introduced a noel problem for analyse members' web surfing data to summarize the fine-grained knowledge acquired by them. A two-step framework is proposed for mining fine-grained knowledge: (1) web surfing data is clustered into tasks by a nonparametric generative model. (2) a novel discriminative infinite Hidden Markov Model is developed to mine fine-grained aspects in each task. Finally, the classic expert search method is applied to the mined results to find proper members for knowledge sharing.

REFERENCE

[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, 2006, pp. 43–50.

[2] 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.

[3] 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.

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

[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., 2003, pp. 17–24.

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

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

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

[9] 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.

[10] H. Deng, I. King, and M. R. Lyu, "Formal models for expert find- ing on DBLP bibliography data," in Proc. IEEE 8th Int. Conf. Data Mining, 2009, pp. 163–172.

[11] 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.

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

