A Theme-Context Mixture Model for Personalized Search in Social Network

Dongling Chen
School of Information Engineering
Shenyang University
Shenyang, 110044,China
e-mail: 13504183184@163.com

Zeng Wen
Institute of Scientific and Technical Information
Beijing 100038, China
e-mail: syucdl@msn.com

Abstract—Nowadays, social network technology provided a lot of ways for users to express their emotions and attitudes online. How to model user preferenced information and provide personalized service is a crucial problem in big data era. In this paper, a new probabilistic model be proposed to model and analysis topic trends in personalized search. The model extended the Latent Dirichlet Allocation (LDA) model by introducing context variables, through which we can detect and analysis topic trends according to contextual information. The core idea of proposed probabilistic model is to learn a finite Dirichlet mixture model, and then adopt Bayesian discriminant to detect topic and topic trends analysis. Experimental results show that the proposed probabilistic mixture model can detect topics and discover topic trends effectively.

Keywords-Mixture Mode; Contextual Mining; LDA; PLSA; User Preference Information

I. INTRODUCTION

With the rapid development of social network technologies, people can express their sentiments or attitudes toward their preferred topics. How to provide personalized service for user is an outstanding problem.

Nowadays, there are many ways to collect user preference information through social network[1,2]. Generally analyze user preferred document can predict user preference topic trends but those documents usually associated with various kinds of context information such as time, location, social background, economic background, and so on. For example, some documents written in the period of some major event ,so it must be influenced by the event in some way. So, it is necessary to consider context information when analyzing the topics covered in such data. Indeed, there have been many studies on this direction. For example, the time stamps of text documents have been considered in some recent work on temporal text mining [3, 4, 5, 6, 7]. Also, author-topic analysis is studied in [8], and cross collection comparative text mining is studied in [9]. All these studies consider some kinds of context information, i.e., time, authorship, and sub-collection[3].

However, those existing techniques are usually applied to some specific tasks. When new context information every time is added into the existing contextual analysis, this drawback appears outstanding. So, we have to seek for a general solutions to copy with complex contextual case in personalized search.

In this paper, a new general probabilistic model for topic trends analysis be proposed. The method extended the Latent Dirichlet Allocation model[10] by introducing context variables, which follow Dirichlet prior, to model the context of a document. The core idea of proposed probabilistic model is to learn a finite Dirichlet mixture model, and then adopt Bayesian discriminant to detect topic and topic trends analysis. Through learning this model, we can discover the global salient themes; analyze the content variation of the themes in any given view of context.

This paper is organized as follows. In Section 2, discuss three related probabilistic generative model, and introduce the contextual mixture model in Section 3. In Section 4 discussed the Gibbs sampler, in Section 5 present the two results on computer science documents-NIPS papers dataset and abstracts dataset from the CiteSeer database. Discussed further research in Section 6.

II. RELATED WORK

Get underlying user preference from large-scale web information, we can only use latent class model[11], which can easily realize unsupervised learning. The basic form as follows: Let y denote discrete dependent, output variable, and z denote a vector of independent, input predictor variables. Then unsupervised latent class can be described:

$$p(y,z) = p(y)p(z \mid y)$$
$$= p(y)\sum_{x} p(x \mid y)p(z \mid y,x) \quad (1)$$

There is a number of recent approaches accordance with this form: PLSA[12,20], LDA[10,22,23], Author-Topic model [13,14], Author-Recipient-Topic model [14,15,17].

A. Probabilistic Latent Semantic Analysis

The probabilistic latent semantic analysis model (PLSA) proposed by Hofmann [2,3,18,19] models a document as a mixture of aspects, where each aspect is represented by a multinomial distribution over the whole vocabulary. Thus, each word is generated from a single topic, and different words in a document may be generated from different topics. Each document is represented as a list of mixing proportions for these mixture components and thereby reduces to a probability distribution on a fixed set of topics. There are about three steps in the process of generating documents. Firstly, generate a topic mixture distribution according to probability P(.|*d*) for each

document. Secondly, select a latent topic z according to probability P($z|d$) for each word. Finally, generate the word according to probability P($w|z$). The document d is generated with probability as follows:

$$p(w_1,...w_N)=\prod_{i=1}^{N}\sum_{z}p(w_i\mid z)p(z\mid d) \qquad (2)$$

But in PLSA, each document is represented as a list of the mixing proportions for topics, and there is no generative probabilistic model for these numbers [4]. This leads to several problems: (1) the number of parameters in the model grows linearly with the size of the corpus, which leads to serious problems with overfitting. (2) It is not clear how to assign probability to a document outside of the training set.

B. Latent Dirichlet Allocation Model

LDA Model proposed by Blei and co-authors[4], which can avoid PLSA shortcomings above mentioned and extract a set of themes from a document collection. A document is considered as a mixture of topics. Each topic corresponds to a multinomial distribution over the vocabulary. The existence of observed word w in document *d* is considered to be drawn from the word distribution Φ_z, which is specific to topic z. Similarly, the topic *z* was drawn from the document-specific topic distribution θ_d.

The LDA model overcome the shortcomings of PLSA, however, this topic model has some shortcomings, for example, someone think it cannot provide explicit information about the interests of authors, who may write several documents often with co-authors and it is consequently unclear how the topics used in these documents might be used to describe the interests of the authors. So, LDA model has been extended step by step by Mark Steyvers and his co-authors [13,14], and Andrew McCallum, and his co-authors[15], respectively generated the Author-Topic model [13, 14] and Author-Recipient-Topic model.

C. Author Topic Model

Author-Topic (AT) model is proposed in [13, 14]. AT model is a similar Bayesian network, in which each authors' interests are modeled with a mixture of topics. In its generative process for each document, a set of authors is observed. To generate each word, an author x is chosen at uniform from this set, then a topic z is selected from a topic distribution θx that is specific to the author, and then a word w is generated by sampling from a topic-specific multinomial distribution Φ_z.

D. Author-recipient-Topic Model

The ART model is a Bayesian network that simultaneously models message content, as well as the directed social network in which the messages are sent. In its generative process for each message, an author, ad, and a set of recipients, rd, are observed. To generate each word, a recipient, x, is chosen at uniform from rd, and then a topic z is chosen from a multinomial topic distribution θ, where the distribution is specific to the author-recipient pair (*ad*, *x*). This distribution over topics could also be smoothed against a distribution conditioned on the author only, although we did not find that to be necessary in our experiments. Finally, the word w is generated by sampling from a topic-specific multinomial distribution Φ_z. The result is that the discovery of topics is guided by the social network in which the collection of message text was generated [15].

In conclusion, there are two basic kinds of probabilistic models about latent factors mining can used for personalized search, one is PLSA model, and another is LDA family models. Those models, especially LDA family models, are all can be straightforwardly applied on large-scale, dynamic topic analysis in which the dimension of topic may increase or decrease as time goes by.

However, an outstanding difficulty in mixture analysis of those approaches is choosing the number of mixture components. They usually treat the number of components as an unknown constant and set its value based on the observed data. Such an approach seems to be too limited, because we wish to model the all the things not only observations but also unseen components. So they don't fit for modeling user preference in personalized search. Especially, when new context information every time is added into the existing contextual analysis, this drawback appears to be outstanding. So, we have to seek for a general solutions to copy with complex contextual case in web2.0, especially, in web2.0 personalized search.

Our work is integrating all those model merits then provides new probabilistic generative model to model user preference information by extended LDA model. We introduce a contextual parameter in LDA model, through different contextual view analysis we can get user preferred topic trends. In addition, we incorporate the Dynamic Topic Model [16,20,21], give the dynamic topic trends analysis which can be applied to all kinds of user preference analysis cases. Also, in our framework we overcome the assumption that one document only belongs to one topic.

III. THEME-CONTEXTUAL MIXTURE MODEL

In this paper, a Theme-Context Mixture Model be proposed, TCMM is extended to LDA. Detailed TCMM in following.

A. Data Space Description

In Personalized search, user preferred information analysis proved effective. Especially, facing dazzled web2.0, actor has enough authority to express his own opinion or revise other actors' opinion on something. But researchers often find themselves in an embarrassing situation, in practical case, large-scale, no or semi- format data, multiple themes, complex contextual make us embarrassed. Existing techniques fail to copy with those problems, although some researcher have proposed some method to copy with this complex case, such as PLSA [8,20], LDA [10,21,22], theme topic [16,17], but they cannot copy with complex contextual analysis. Even though in [3], Mei Qiaozhu proposed CPLSA, it is PLSA extended model, so CPLSA inherit PLSA shortcomings. So, it is highly desirable to introduce a general text contextual mining, which is abstracted from a family of text mining tasks with various types of contextual analysis [3].

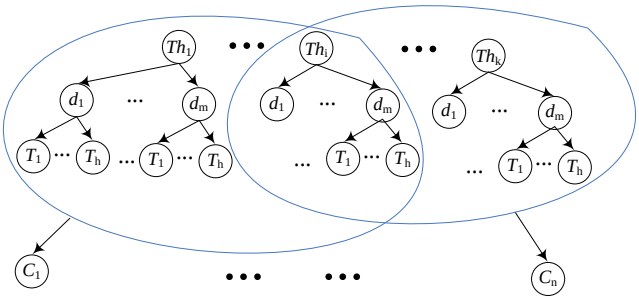

Figure 1. Document Collection Graphical Representation.

In this section, we give a general analysis on contextual text analysis. Given a large-scale text collections, it contain many documents, which may contain multiple topics and complex contextual content. In addition, those documents maybe give an expression for some themes of document collection. So, we give a graphical presentation as Fig .1 show.

Here, Th1-Thk denotes theme parameters, d1-dm denotes document parameters, T_1-Thdenotes topic parameters, C_1-C_n denotes contextual parameters. Accordance with Fig .1, in analysis process, we must define following parameters: theme, document, topic, context, and presentation. Among those parameters, presentation parameter denotes generate a final view on user preferred topic.

B. *Theme-Context Mixture Model*

1) *Formally Description*

First, we give several variables, and then give some important relevant definitions. In this model the observed variable is the context variable cdi, $i \in \{1...n\}$ and variable x is a multinomial distribution over cdi, it is the mixing proportions of sampled context content. The unseen variable is the theme variables Th_j, Th_j~Dir(α), $j \in \{1,...,k\}$ and variable θ_l are proportions sampling from Th, it contain two parts: $\{(Th_{l1},\omega_1),...,(Th_{ls},\omega_s)\}$, $l \in \{1,...,s\}$, s maybe equal to k ; ω_i is proportion of Th_{li} in θ_l . Th variable z is a topic variable, $t \in \{1,...,m\}$, z is sampled from pair variable Ω, it is pair(context, Th_i, proportioni) distribution, which follow Dirichlet distribution, i.e., Ω~Dir(γ). In addition, D, Nd, T, J and pair (content, Th_i, proportioni) numbers are hyper-parameters that must be chosen. Fig .1 gives the graphical representation of TCMM.

In the next we give several relevant definitions:

Definition1. Context: let F={f1,…,f|F|} be a set of context features, which can be captured by relevance feedback in personalized search.(for example, some political information or economic information in user click-through data) A context C in a document collection is decided by any combination of preference features in F, The whole set of possible context is denoted as C={c1,…,cn}. a document can belong to multiple contexts, in another word; the contexts maybe overlap.

Definition2. Theme: in large-scale web corpus, some public topic is theme. It is obviously that we often cluster some similar data based on themes. Th={Th_1,…Th_k}, Th_i maybe consist of one or more topics. By integrated themes and contexts, we can observe the growth or decay of each theme mixing component to discover user preferred topic trends. A document may consist of several topics.

Definition3. Document Topic Coverage: Document topic coverage is the coverage of themes in a document (Th_j) i.e., it is a probabilistic distribution over the themes $p(l \mid Th_i)$ clearly, $\sum_{l=1}^{k} p(l \mid Th_i) = 1$

Definition4. Document Theme: a document maybe belong to multiple contexts and multiple themes, in model, we design a tri-tuples (C_i, theme, theme proportion) to describe different theme in a document. Lay aside context, only thinking the theme and its proportion, we are concerned with the top k theme, then characterized this document by those top-k themes, respectively.

2) *Generative Model and Parameters Estimation*

The TCMM model is a Bayesian network that simultaneously models contextual content, as well as the theme proportions. In its generative process for large-scale text corpus user preferred, context content, C_i, are observed, the theme are unseen and Th~Dir(α). x, denotes sampling from contextual content by random proportions, i.e., each context is characterized by a particular value for the mixing proportions over x. the variable θ denotes select one theme from corpus space. To generate each word, a theme θ is chosen from Th, and then, x, a multinomial distribution over C_d, is to form a multinomial distribution $\boldsymbol{z}$, where the distribution is specific to the theme-context pair (θ, x). Finally, the word w is generated by sampling from a topic-specific multinomial distribution z. The result is that the corpus was generated by the mixture discovery of (context, theme, theme proportions) pairs. i.e., the basic idea is that documents are represented as random mixtures over context and pair (theme, theme proportions). The graphical representation of Theme-Context mixture model is shown in Fig .2 .

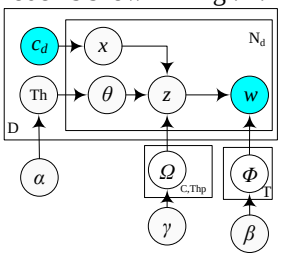

Figure 2. A graphical representation of Theme-Context mixture model

In the TCMM model, for a particular document d, given the hyper-parameters α, β, γ, the context content cd, and the set of themes Th, the joint probability distribution as follows:

$$p(x,\theta,z,\Omega,\phi,w,Th \mid c_d,\alpha,\beta,\gamma) = p(Th \mid \alpha)\,p(\Omega \mid \gamma)\,p(\phi \mid \beta)\prod_{n=1}^{n_d} p(x_n \mid c_d)\,p(\theta_n \mid Th)\,p(z_n \mid \Omega)\,p(w_n \mid \phi) \quad (3)$$

Integrating over Th, Ω, Φ and summing over x, θ, z, we obtain the marginal distribution of a document: To calculate marginal distribution of document, as follows:

$$p(w \mid c_d,\alpha,\beta,\gamma) = \iiint p(Th \mid \alpha)\,p(\Omega \mid \gamma)\,p(\phi \mid \beta)\prod_{n=1}^{n_d}\sum_{x}\sum_{\theta}\sum_{z} p(x_n \mid c_d)\,p(\theta_n \mid Th)\,p(z_n \mid \Omega)\,p(w_n \mid \phi)\,dTh\,d\Omega\,d\phi \quad (4)$$

Finally, the corpus can be generated as follows:

$$P(D) = \prod_{d=1}^{D} p(w \mid c_d,\alpha,\beta,\gamma) \quad (5)$$

Three standard approximations have been used to obtain practical results: variational methods [10], Gibbs sampling [12, 13, 14], and expectation propagation [11, 12]. We chose Gibbs sampling for its ease of implementation.

IV. EXPERIMENT AND EVALUATION

In order to evaluate the effectiveness of the proposed PLSA based constructing user profile approach, and better improve the effect of personalized search, we conducted some preliminary experiments and made comparisons with actual user interests well.

A. Data Sets

We invited six volunteers, and search engine we selected is Google. We implicitly monitored their searching and browsing activities through proxy for approximately consecutive four weeks in order to avoid that data is too sparse. We built two data sets according to the logged data:

First, we collected about 102 blog articles, 400 queries in all submitted by the users and 2496 web pages they selected to browse from the results given by the Google. For each user, we also divided his sessions into 17 subsets, among them, 14 as training sessions，those used to create profiles; and 3 as testing sessions, those used to evaluate the effectiveness of user profiles.

Secondly, according to time sequences, we divided the sessions of per user into two subsets. The first three weeks sessions are selected as training set, while the remainder as testing set. Meanwhile, we make preprocess to those data, we remove manually the sessions which have little content-related between training set and testing set. The purpose of building this data set is to make comparisons between the personalized search results based on user profile and actual user interests.

Table 1 gives the statistics of the data sets. For example, user 1 has 10 interest categories, and 64 queries and 420 relevant documents.

B. Performance Evaluation

Accuracy of Mapping User Queries to Categories

In our study, after we built an initial user profile, the most important thing next to do is the user profile's maintenance and updating according to algorithm3. In order to evaluate the accuracy of algorithm3, we should propose a metric to evaluate the accuracy of mapping a user query to a set of categories. Due to all the queries was mapped into categories, which are used to appended or updating the user profile, the utility of those categories is a hundred percent, so from the algorithm3 per se, we can't find a method to solve this problem. We only fall back on the user's feedback at this point. Finally, the metric is proposed: $Accuracy = \frac{m}{n}$, where m is the number of related categories to the query, which is assigned correct certainly by user, while n is the number of all the categories. According to the user's feedback, the accuracy usually is about 70%.

TABLE I. STATISTICS OF THE DATA SETS

Statistics	User 1	User 2	User 3	User 4	User 5	User 6
# of Interest categories	10	6	4	5	6	8
# of Queries	64	56	46	66	88	80
# of Sessions	17	17	17	17	17	17
# of Web pages	420	385	430	398	450	413
Average #of queries in one sessions	3.7	3.3	2.7	3.9	5.2	4.7

C. Comparison with Actual User Interests

Although accuracy of algorithm3 is measured, but the whole accuracy of the generated user profiles doesn't be measured. Thus, we still fall back on user to feedback to evaluate how appropriately these user profiles reflected their interests. We want to know:

1.How many categories in their profiles do reflect their actual interests?

2. How well do those categories reflect their actual interest?

3. How well does the entire user profile describe your actual interests?

4. Whether the learning and updating algorithm can improve the degree of user profile reflects their actual interests

Experimental results proved that the performance of the user profile is exciting. When the number of the categories is 10 in user profile, we list the results of average user feedback as follows:

TABLE II. USER'S FEEDBACK

How well those categories reflect your actual interests	8
How many categories can reflect your interests	85%
How well entire user profile reflect your actual interests	80%
Whether the learning algorithm improved performance	yes

Also, experimental results indicate that as more training sessions are given, the accuracy of user profile reflect the user's actual interests increases.

According to Probability Latent Semantic Analysis model, we construct user profile, consequently, better perform personalized search. In the procedure of constructing user profile, we utilize Bayes Probability equation to calculate the latent factor under the co-occurrence data, in terms of those underlying user search intention, we built user profiles. Our approach avoids shortcomings which other constructing user profile approach owned. Also, we proposed an effective algorithm to learning and updating user profiles.

V. CONCLUSION AND FUTURE WORK

In this paper, we have presented an overview of several density estimation models for document representation. We have furthermore proposed yet another model in this family, namely the Theme Topic Mixture Model (TTMM), which lies in between LDA and PLSA, sharing some advantages of both of them. A theoretical comparison between the models was then presented, highlighting advantages and problems of each method.

TABLE III. LEARNING TABLE

session	user 1	user 2	user 3	user 4	user 5	user 6
1	0.31	0.342	0.335	0.281	0.321	0.346
2	0.38	0.573	0.401	0.398	0.325	0.365
3	0.453	0.567	0.408	0.405	0.452	0.452
4	0.467	0.773	0.658	0.407	0.462	0.453
5	0.485	0.768	0.628	0.407	0.48	0.512
6	0.488	0.763	0.677	0.407	0.492	0.564
7	0.483	0.81	0.804	0.594	0.512	0.654
8	0.479	0.821	0.881	0.598	0.526	0.687
9	0.569	0.817	0.884	0.601	0.661	0.701
10	0.671	0.88	0.884	0.604	0.668	0.714
11	0.688	0.865	0.887	0.621	0.721	0.744
12	0.693	0.89	0.889	0.633	0.753	0.768
13	0.75	0.891	0.89	0.694	0.785	0.798
14	0.879	0.897	0.89	0.694	0.798	0.84

This was followed by an empirical analysis, which shows that no one model is always better than the others, and that the ultimate choice may depends on the actual data configuration. Interestingly, all of the proposed density estimation models fail with respect to the simple bag-of-words representation when the size of the dataset become large. This probably means that the constraints that have been purposely integrated into all these models (the choice of words in a document is independent of the document itself given a hidden topic variable) may be useful when the data is scarce but too strong when it is abundant, in which case constraints should be relaxed somehow.

ACKNOWLEDGMENTS

This work is supported by Liaoning Province PhD research start up fund. (No.20101074), National Social Science Fund of China.(No. 14BTQ038）

REFERENCES

[1] Chen D.L., Wang D.L., Yu G.: A PLSA-Based Approach for Building User Profile and Implementing Personalized Recommendation. Springer-Verlag, Berlin Heidelberg New York, In APWeb/WAIM'07: 606-613

[2] http://lrs.ed.uiuc.edu/tse-portal/analysis/social-network-analysis/, November, 2001

[3] Mei.Q. Zhai C.X. A Mixture model for contextual text mining. In KDD'06 :649-655

[4] J. Kleinberg. Bursty and hierarchical structure in streams. In Proceedings of KDD '02, pages 91‑101.

[5] J. Perkio, W. Buntine, and S. Perttu. Exploring independent trends in a topic-based search engine. In Proceedings of WI '04, pages 664‑668, 2004.

[6] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In Proceeding of KDD'05, pages 198‑207, 2005.

[7] C. C. Chen, M. C. Chen, and M.-S. Chen. Liped: Hmm-based life profiles for adaptive event detection. In Proceeding of KDD '05, pages 556‑561, 2005.

[8] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In Proceedings of KDD'04, pages 306‑315, 2004.

[9] C. Zhai, A. Velivelli, and B. Yu. A cross-collection mixture model for comparative text mining. In Proceedings of KDD'04, pages 743–748, 2004.

[10] D. M.Blei, A.Ng, and M. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, Vol3:pp993-1022, 2003.

[11] T. L. Griffiths, and M. Steyvers (2004). Finding scientific_topics. Proceedings of the National Academy of Sciences, 101 (suppl. 1), 5228-5235.

[12] M. Rosen-Zvi, T. Griffiths, M. Steyvers, P.Smyth. The Author-Topic Model for Authors and Documents, In 20th Conference on Uncertainty in Artificial Intelligence. Banff, Canada

[13] M. Steyvers, P.Smyth., T. Griffiths. Probabilistic Author-Topic Models for Information Discovery. In:10th ACM sigKDD conference knowledge discovery and data mining.2004.

[14] A.McCallum, A. Corrada-Emmanuel, X.Wang. The Author-Recipient-Topic model for topic and role discovery in social networks: experiments with enron and academic email. Technical Report UM-CS-2004-096.

[15] D.M. Blei, J.D. Lafferty. Dynamic Topic Models. in Proceedings of the 23 rd International Conference on Machine Learning, Pittsburgh, pp.113-120, 2006.

[16] M.Keller, S. Bengio, Theme topic mixture model for document representation. in PASCAL Workshop on Learning Methods for Text Understanding and Mining, 2004.

[17] S. Basu, M. Bilenko, and R.J. Mooney. A probabilistic framework for semi-supervised clustering. In Proc. of KDD2004.

[18] T. Hofmann. Probabilistic Latent Semantic Analysis. The 22nd Annual ACM Conference on Research and Development in Information Retrieval, Berkeley, California: ACM Press, 1999, pp 50-57.

[19] Q.Z. Mei, X.Ling, M.Wondra, et,al. Topic Sentiment Mixture: Modeling facets and opinions in weblogs. Canada: ACM Press. WWW 2007.

[20] Y. W. Teh, M. I. Jordan, M. J. Beal, & D. M. Blei. Hierarchical dirichlet processes(Technical Report). UC Berkeley Statistics.2004.

[21] Y. Yu, Q.F. Xu, P.F. Sun, Bayesian clulstering based on fiinite mixture models of dirichlet distribution.[J] MATHEMATICA APPLICATA. 2006, 19(3):pp600-605.