A Mixture Model for Contextual Text Mining

Qiaozhu Mei
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana,IL 61801
qmei2@uiuc.edu

ChengXiang Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana,IL 61801
czhai@cs.uiuc.edu

ABSTRACT

Contextual text mining is concerned with extracting topical themes from a text collection with context information (e.g., time and location) and comparing/analyzing the variations of themes over different contexts. Since the topics covered in a document are usually related to the context of the document, analyzing topical themes within context can potentially reveal many interesting theme patterns. In this paper, we propose a new general probabilistic model for contextual text mining that can cover several existing models as special cases. Specifically, we extend the probabilistic latent semantic analysis (PLSA) model by introducing context variables to model the context of a document. The proposed mixture model, called contextual probabilistic latent semantic analysis (CPLSA) model, can be applied to many interesting mining tasks, such as temporal text mining, spatiotemporal text mining, author-topic analysis, and cross-collection comparative analysis. Empirical experiments show that the proposed mixture model can discover themes and their contextual variations effectively.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Text Mining

General Terms: Algorithms

Keywords: Contextual text mining, context, mixture model, EM algorithm, theme pattern, clustering

1. INTRODUCTION

A text document is often associated with various kinds of context information, such as the time and location at which the document was produced, the author(s) who wrote the document, and its publisher. The contents of text documents with the same or similar context are often correlated in some way. For example, news articles written in the period of some major event all tend to be influenced by the event in some way, and papers written by the same researcher tend to share similar topics. In order to reveal interesting content patterns in such contextualized text data, it is necessary to consider context information when analyzing the topics covered in such data. Indeed, there have been several recent studies in this direction. For example, the time stamps of text documents have been considered in some recent work on temporal text mining [9, 16, 14, 4]. Also, author-topic analysis is studied in [17], and cross-collection comparative text mining is studied in [18]. All these studies consider some kinds of context information, i.e., time, authorship, and subcollection.

Time, authorship, and subcollection are by no means the only possible context information of a document. In fact, any metadata entry of a document can indicate a context and all documents with the same value of this metadata entry can be considered as in the same context. For example, the source of a news article, the author's age group, occupation, and location of a weblog article, and the citation frequency of a research paper, are all reasonable context information. Moreover, a document may belong to multiple contexts, and any combination of its metadata entries makes a "complex" context. By analyzing the variations of topics over these contexts, a lot of interesting text mining tasks can be addressed, such as spatiotemporal text mining, author-topic evolutionary analysis over time, and opinion comparison over different age groups and occupations.

However, existing techniques are usually tuned for some specific tasks, and are not applicable to consider other kinds of contexts. For example, one cannot directly use the temporal text mining techniques to model the occupation of authors. This indicates a serious limitation of existing contextual analysis of themes: every time when a new combination of context information is to be considered, people have to seek for solutions in an ad hoc way.

Therefore, it is highly desirable to introduce a general text mining problem, contextual text mining, which is abstracted from a family of text mining tasks with various types of contextual analysis. It is desirable to derive a model that is highly general to conduct the common tasks of these specific contextual text mining problems, and easy to be applied to each of them with appropriate regularization.

In this work, we define the general problem of **Contextual Text Mining (CtxTM)** and its common tasks, which is abstracted from a family of specific text mining problems. We extend the probabilistic latent semantic analysis (PLSA) model to incorporate context information, and develop a contextual probabilistic latent semantic analysis (CPLSA) model to facilitate contextual text mining in a general way. By fitting the model to the text data to mine, we can (1) discover the global salient themes from the collection of documents; (2) analyze the content variation of the themes in

KDD'06, August 20–23, 2006, Philadelphia, Pennsylvania, USA.

any given view of context; and (3) analyze the coverage of themes associated with any given context.

These tasks are general and can be easily applied to different specific contextual text mining problems. In this paper, we show that many existing contextual theme analysis problems can be defined as special cases of **CtxTM**, and can be solved with regularized versions of the mixture model we proposed, corresponding to the context information and the mining tasks it involves. Although it may not be the only possible model for contextual text mining, the model is quite flexible to adapt different assumptions.

2. CONTEXTUAL TEXT MINING

Given a collection of documents with context information, we assume that there is a set of topics, or themes in the collection which vary over different contexts. Our goal is generally to conduct context-sensitive analysis of these themes. As stressed in previous work [14], a ***theme*** in a contextualized text collection D is a probabilistic distribution of words that characterizes a semantically coherent topic or subtopic. Without loss of generality, we will assume that there are altogether k major themes in our collection, $\Theta = \{\theta_1,..., \theta_k\}$.

To model the context of a document, we introduce a concept called *context feature*, which is defined as any meta-data of a document (e.g., the *time stamp* in temporal text mining or *authorship* in author-topic analysis). The *context* that a document belongs to can be indicated by the context features of this document, which is formally defined as follows:

Definition 1 (Context) Let $\mathcal{F} = \{f_1, f_2, ..., f_{|F|}\}$ be a set of context features. A ***Context*** c in a document collection is decided by any combination of context features in $\mathcal{F}$, formally $c \subseteq 2^{|\mathcal{F}|}$. The whole set of possible contexts is denoted as $\mathcal{C} = \{c_1, ..., c_n\}$. Suppose $\mathcal{D} = \{(D_1, C_1), ..., (D_{|\mathcal{D}|}, C_{|\mathcal{D}|})\}$ is a collection of documents, each document D_i is a sequence of words from a vocabulary set $V = \{w_1, ..., w_{|V|}\}$, and $C_i \subseteq \mathcal{F}$ is a set of context features which are associated with the document D_i. A document D_i belongs to a context c iff. $C_i \subseteq c$. This tells us that a document can belong to multiple contexts. In another word, the contexts are possible to overlap.

In contextual text mining, our goal is to analyze the topics/subtopics in such a text collection in a context-sensitive way. Specifically, we would like to model the k major themes and how they vary according to different contexts, and would also like to model the coverage of different themes in a document or documents that share certain kinds of context.

To accommodate context-sensitive theme analysis, we consider variations of these k themes over different contexts. For example, if the context we are interested in is time, we will assume that there is a potentially distinct "version" of the k themes in each different time period; different such "versions" model the variations of themes across time stamps. We formally define such a variation as a *View* of themes.

Definition 2 (View) A ***view*** of themes in a contextualized text collection D is a sequence of themes $\theta_{i1}, ..., \theta_{ik}$, where θ_{il} is the variation of theme θ_l according to view v_i.

We will assume that there are n views in our collection, $v_1, ..., v_n$, each corresponds to a context c_i. Therefore, a document is assumed to potentially have multiple views; precisely which views are taken depends on the document and its context. Each view v_i is assumed to be taken in any documents in the context c_i, which can also be overlapping.

Definition 3(Context Support) The ***support*** of a context c_i, $s(c_i)$ is the set of documents in context c_i, i.e., $s(c_i) = \{D_j | C_j \subseteq c_i\}$. Since each context is associated with a view, we also call $s(c_i)$ as the ***support*** of the view v_i.

To analyze the strength of themes, we further model the variable coverage of different themes in a document. For example, some documents would favor some particular themes and thus would have a larger coverage of them.

Definition 4 (Coverage) A ***coverage*** of themes in a document (κ_j) is a distribution over the themes $p(l|\kappa_i)$. Clearly, $\sum_{l=1}^{k} p(l|\kappa_j) = 1$.

We will assume that there are m distinct theme coverages in our collection, κ_1, ..., κ_m. For example, if we assume that each document has a potentially distinct theme coverage, then $m = |\mathcal{D}|$. In general, however, a document can cover themes according to multiple coverages. For example, if we are interested in modeling theme coverage associated with time stamps, we may assume that the actual theme coverage in a document would be a mixture of the document-specific theme coverage and another theme coverage associated with the time context of the document.

We use $c(\kappa_j)$ to denote the contexts where the coverage κ_j is applicable, and we also define the support of a coverage in the same way as we define that of a context.

Definition 5 (Coverage Support) The ***support*** of a coverage κ_j, $s(\kappa_j)$ is the set of documents in which the coverage κ_j is taken, i.e., $s(\kappa_j) = \{D_i | \exists c \in c(\kappa_j)\ s.t.\ C_i \subseteq c\}$.

The latent structure of themes, views and coverages in a contextualized document collection is illustrated in Figure 1.

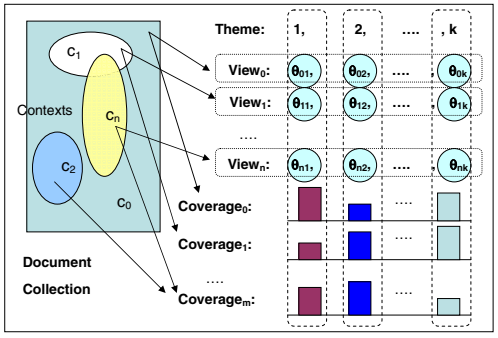


Figure 1: The theme-view-coverage structure in a text collection

With these definitions, the task of **Contextual Text Mining (CtxTM)** can be defined as to recover the n views, $v_i = (\theta_{i1}, ..., \theta_{ik})$, $i = 1, ..., n$, and the m theme coverages $\kappa_1, ..., \kappa_m$ from the collection $\mathcal{D}$, and to analyze them in a context-sensitive way. There are many different ways to analyze the views and theme coverages. Below we discuss a few interesting cases.

1. Theme extraction: We may extract the global salient themes. Although each theme θ_l varies in different contexts, it is also beneficial to have an explicit model for θ_l in a global view. Basically, this will give us the common information that is shared by all the variations of θ_l in all different contexts. In practice, we always include a global view v_0, which corresponds to a global context $c_0 = \mathcal{F}$. Clearly, all documents $D_i \in \mathcal{D}$ belong to c_0 since $C_i \subseteq c_0$.

2. View comparison: We may compare the n views. The comparison of a theme θ_l from different views usually represents the content variation of θ_l corresponding to dif-

ferent contexts. By comparing θ_{il} for each view v_i which corresponds to context c_i, we can analyze the influence of the context c_i on the contents of θ_l.

3. Coverage comparison: We may compare the m coverages. The variations of $p(l|\kappa_j)$ can tell us how likely θ_l is covered by the documents in the coverage support $s(\kappa_j)$. By associating $p(l|\kappa_j)$ within contexts $c(\kappa_j)$ that κ_j is applicable, we can analyze how closely a theme is associated to a context, or how context-sensitive a theme is.

4. Others: With contextual text mining, we can also analyze other problems such as the influence of an individual context feature on the theme coverage, e.g., the theme-location distribution in spatiotemporal theme analysis.

Among these cases, 2 and 3 are the most important, which distinguish contextual text mining from the traditional theme extraction work, and the application of them facilitate other types of analysis.

With the definition of the general problem of contextual text mining (CtxTM), we can show that some specific contextual text mining problems are special cases of CtxTM. For example, in temporal text mining, each context feature is a time stamp. Therefore, a context is either a time stamp or a set of consecutive time stamps, or time period. A view of themes is taken in all the documents in the corresponding time. The goal of temporal text mining is mainly to compare the coverage variation over different contexts (e.g., theme life cycles in [14]), and sometimes also the content variation of themes over different views, (e.g., evolutionary theme pattern in [14]). In author-topic analysis, each context feature is an author, and each context is either an author or a set of authors. Each view is then taken in the document with the same author or authors. We are interested in comparing the content variations over different views (authors) [17].

3. A CONTEXTUAL MIXTURE MODEL

In this section, we propose an extension of the Probabilistic Latent Semantic Analysis (PLSA) model [7, 8], called Contextual Probabilistic Latent Semantic Analysis (CPLSA) model, for contextual text mining. Our main idea is to allow a document to be generated using multiple views and multiple coverages. The views and coverages actually used in a document usually depend on its context, which could be the time or location where the document is written, the source from which the document comes, or any other metadata. We first propose the general CPLSA model, and then introduce two simplified versions of this model that are especially suitable for two representative tasks of contextual text mining.

3.1 The CPLSA Model

In CPLSA, we assume that document D (with context C) is generated by generating each word in it as follows: (1) Choose a view v_i according to the view distribution $p(v_i|D,C)$. (2) Choose a coverage κ_j according to the coverage distribution $p(\kappa_j|D,C)$. (3) Generate a word using θ_{il}.

Formally, the log-likelihood of the whole collection is

$$\log p(\mathcal{D}) = \sum_{(D,C)\in\mathcal{D}}\sum_{w\in\mathcal{V}} c(w,D)\log(\sum_{i=1}^{n} p(v_i|D,C) \times \sum_{j=1}^{m} p(\kappa_j|D,C)\sum_{l=1}^{k} p(l|\kappa_j)p(w|\theta_{il}))$$

The parameters are the view selection probability $p(v_i|D,C)$, the coverage distribution selection probability $p(\kappa_j|D,C)$, the coverage distribution $p(l|\kappa_j)$, and the theme distribution $p(w|\theta_{il})$.

As a mixture model, we have a total of $n \times k$ multinomial distribution component models. Each set of k multinomial distributions, $\theta_{i1}, ..., \theta_{ik}$, represents a potentially distinct view of the topics that we are interested in. However, while we can potentially use all the views to generate a document, often the generation of a particular document D in a particular context C only involves a subset of these views. This is because in any interesting context mining scenario, different views generally have different supporting documents, though it is also common for the views to overlap in some supporting documents.

More specifically, the view selection distribution $p(v_i|D,C)$ determines which views will actually be used when generating words in document D. This distribution would assign zero probabilities to those views that are not selected. For example, if the views that we are to model correspond to the temporal context of a document and we have one global view spanning in the entire time period, then a document at time point t_i would be generated using two different views – the view corresponding to time point t_i and the global view, which is applied to all the documents.

Orthogonal to the choice of views, we also assume that we have choices of theme coverage distributions. The different coverage distributions are to reflect the uneven coverage of topics in different context and to capture the common coverage patterns. For example, if we suspect that the coverage may vary depending on the location of the authors, we can associate a particular coverage distribution to each location, which will be shared by all the documents in the location. After we learn such coverage distributions, we can then compare them across different locations. Once again, exactly which coverage distributions to use would depend on the context of the document to be generated.

The mixture model can be fit to a contextualized collection $\mathcal{D}$ using a maximum likelihood estimator. The EM algorithm [5] can be used in a straightforward way to estimate the parameters; the updating formulas are as follows:

$$p(z_{w,i,j,l}=1) = \frac{p^{(t)}(v_i|D,C)p^{(t)}(\kappa_j|D,C)p^{(t)}(l|\kappa_j)p^{(t)}(w|\theta_{il})}{\sum_{i'=1}^{n} p^{(t)}(v_{i'}|D,C)\sum_{j'=1}^{m} p^{(t)}(\kappa_{j'}|D,C)\sum_{l'=1}^{k} p^{(t)}(l'|\kappa_{j'})p^{(t)}(w|\theta_{i'l'})}$$

$$p^{(t+1)}(v_i|D,C) = \frac{\sum_{w\in V} c(w,D)\sum_{j=1}^{m}\sum_{l=1}^{k} p(z_{w,i,j,l}=1)}{\sum_{i'=1}^{n}\sum_{w\in V} c(w,D)\sum_{j=1}^{m}\sum_{l=1}^{k} p(z_{w,i',j,l}=1)}$$

$$p^{(t+1)}(\kappa_j|D,C) = \frac{\sum_{w\in V} c(w,D)\sum_{i=1}^{n}\sum_{l=1}^{k} p(z_{w,i,j,l}=1)}{\sum_{j'=1}^{m}\sum_{w\in V} c(w,D)\sum_{i=1}^{n}\sum_{l=1}^{k} p(z_{w,i,j',l}=1)}$$

$$p^{(t+1)}(l|\kappa_j) = \frac{\sum_{(D,C)\in\mathcal{D}}\sum_{w\in V} c(w,D)\sum_{i=1}^{n} p(z_{w,i,j,l}=1)}{\sum_{l'=1}^{k}\sum_{(D,C)\in\mathcal{D}}\sum_{w\in V} c(w,D)\sum_{i=1}^{n} p(z_{w,i,j,l'}=1)}$$

$$p^{(t+1)}(w|\theta_{il}) = \frac{\sum_{(D,C)\in\mathcal{D}} c(w,D)\sum_{j=1}^{m} p(z_{w,i,j,l}=1)}{\sum_{w'\in V}\sum_{(D,C)\in\mathcal{D}} c(w',D)\sum_{j=1}^{m} p(z_{w',i,j,l}=1)}$$

However, since the model has many parameters and has a high-degree of freedom, fitting it with a maximum likelihood estimator, in general, would face a serious problem of multiple local maxima. Fortunately, in contextual text mining, we almost always associate them with appropriate partitions of context. As a result, the model is often highly constrained. For example, if all we are interested in is to compare non-overlapping views across different time, then $p(\kappa_j|D,C)$ becomes a delta function, i.e., $p(\kappa_j|D,C)=1$ if and only if κ_j is the coverage distribution for the time context of D, and $p(\kappa_j|D,C)=0$ for all other κ_j's.

Unfortunately, even with such constraints, the model may

still have many free parameters to estimate. One possibility is to add some parametric constraint such as assuming all coverage distributions are from the same Dirichlet distribution as done in LDA [2], which would clearly reduce the number of free parameters; indeed, we can easily generalize our model in the same way as LDA generalizes PLSA [8]. However, one concern with such a strategy is that the parametric constraint is artificial and may restrict the capacity of the model to extract discriminative themes, which is our goal in contextual text mining. Another approach is to further regularize the estimation of the model by heuristically searching for a good initial point in EM; specific heuristics would depend on the particular contextual text mining task. This approach is adopted in our experiments and will be further discussed in Section 3.2.

In order to model the noise (e.g., common English words) in the text, we could designate the first theme as modeling such noise. That is, all θ_{1j}'s will be set to model the noise. We may further tie all of them so that we have just one common background unigram language model θ_1. This can also be regarded as applying an infinitely strong prior on the first theme in all views.

3.2 Special Versions of CPLSA

Considering that the most important tasks of contextual text mining are view comparison and theme coverage comparison across contexts, as discussed in Section 2, we introduce two special cases of CPLSA, which are particularly useful to do these two tasks.

We first introduce the special version of CPLSA to facilitate view comparison. In some cases, we are only interested to model the content variation of themes across contexts, e.g., when we are analyzing the theme evolutions over time [14], or comparing the common themes and corresponding specific themes across subcollections [18]. In these cases, we can fairly assume that the theme coverage over contexts is fixed, thus does not depend on the contexts that a document is in. Under this assumption, the $\sum_{j=1}^{m} p(\kappa_j|D,C)$ in the model will be simplified as $\sum_{j=1}^{m} p(\kappa_j|D)$. If we further assume that there is only one coverage κ applicable to each document, the log-likelihood function can be written as

$$\sum_{(D,C)\in\mathcal{D}} \sum_{w\in\mathcal{V}} c(w,D) \log(\sum_{i=1}^{n} p(v_i|D,C) \sum_{l=1}^{k} p(l|\kappa_D) p(w|\theta_{il}))$$

where κ_D is the coverage associated with the document D. We call this simplified version of model as **fixed-coverage contextual mixture model (FC-CPLSA)**. If we have three views, where one is the global view and the other two correspond to subcollections, it will allow us to compare the common themes and specific themes in the two views, as discussed in [18]. If each view corresponds to a time stamp, this model will allow us to analyze the content evolutions of themes over time, as discussed in [14].

In some other cases, we are only interested to model the variation of theme coverage over contexts, e.g., when we are analyzing the life cycles (i.e., strength variations over time) of themes. In these cases, we are not interested in the content variation of local themes, and thus make the assumption that different views of themes are stable. With this assumption, we can simplify the model likelihood as

$$\sum_{(D,C)\in\mathcal{D}} \sum_{w\in\mathcal{V}} c(w,D) \log(\sum_{j=1}^{m} p(\kappa_j|D,C) \sum_{l=1}^{k} p(l|\kappa_j) p(w|\theta_l))$$

where $p(w|\theta_l)$ is the global word distribution of theme l, which does not vary across contexts. We call this simplified model as **fixed-view contextual mixture model (FV-CPLSA)**. If the only context feature is time, we have two types of coverage distributions κ_D and κ_T, where κ_D is the coverage distribution corresponding to each document and κ_T is the theme coverage for each time period. This will allow us to model the theme life cycles, as introduced in [14]. If we have two context features, time and location, and each context is a combination of time stamp and location, we also have two groups of theme coverage distributions, κ_D and κ_{TL}. This will allow us to analyze the spatiotemporal theme distributions in a spatiotemporal text mining framework.

With these two special simplified versions, the CPLSA model can be applied to solve a broad family of text mining problems with contextual analysis.

4. EXPERIMENTS

We apply the general CPLSA model presented in Section 3 to three different datasets and text mining tasks. Empirical results show that this model can model the themes and their variations across different contexts effectively.

4.1 Temporal-Author-Topic analysis

In this experiment, we evaluate the performance of the CPLSA models on author-topic comparative analysis. If two authors have similar research interest, we assume that there is a set of common themes which can be found in their publications. Since different author has different preferences and focuses, the content of these themes will also vary corresponding to each author. Previous work on author-topic analysis only consider the authorship of documents as the context [17]. Intuitively, however, the topics that an author favors also evolute over time. We add another type of context information, i.e., publication time, to test the effectiveness of our model on handling multiple types of contexts.

We collect the abstracts of 282 papers published by two famous Data Mining researchers from ACM Digital library. We split the whole time line into three spans: before the year 1993, from 1993 to 1999, and after the year 1999. This will give us 12 possible views as in Table 1. Since we are not interested in analyzing the coverage variations across contexts (i.e. time and authors), we assume the coverage of themes only depends on documents but not on the contexts.

#Context Features	Views (A and B are two authors)
0	Global View
1	A; B; < 92; $93 \sim 99$; $00 \sim 05$
2	A, < 92; A, $93 \sim 99$; A, $00 \sim 05$ B, < 92; B, $93 \sim 99$; B, $00 \sim 05$

Table 1: Possible Views in Author-Topic Analysis

Therefore, we use the FC-CPLSA model presented in Section 3.2 to model the themes and their views corresponding to different contexts. Our goal is thus to estimate all the parameters in the regularized model, and compare $p(w|\theta_{jl})$ over different view v_j.

To avoid the EM algorithm being trapped in suboptimal local maximums, we need to make associations between each θ_{jl} to its corresponding global view θ_l. We achieve this by selecting a good starting point for the EM algorithm. Specifically, we begin with a prior of a large $p(v_0|D,C)$ to view 0, which is the global view. This ensures us to get the strong signal of global themes instead of local biased themes. In the following iterations, we gradually decay this prior and

Vews:	Global	Author A	Author B	Author A: 2000∼	1993∼1999	2000∼
Author Topic Analysis	pattern 0.110689	project 0.0444375	research 0.0550772	close 0.0805878	rule 0.0616733	index 0.0430914
	frequent 0.040613	itemset 0.0432976	next 0.0308254	pattern 0.072078	distribute 0.0567852	graph 0.0343051
	frequent-pattern 0.0393	intertransaction 0.03072	transition 0.0308254	sequential 0.0462879	researcher 0.0324659	web 0.0306886
	sequential 0.0359059	support 0.0264818	panel 0.0275384	min_support 0.03526	algorithm 0.0217309	gspan 0.0273849
	method 0.0214187	associate 0.0258175	technical 0.0275384	length 0.0315721	over 0.0162951	substructure 0.02005
	pattern-growth 0.02035	frequent 0.0181942	technology 0.0258949	threshold 0.0296533	fdm 0.0227141	gindex 0.016431
	condense 0.0184008	closet 0.0176081	article 0.0154127	frequent 0.0196054	study 0.0116576	bide 0.016431
	increment 0.0138457	apriori 0.0170468	revolution 0.0154127	top-k 0.0176324	scable 0011357.	magnitude 0.0151909
	constraint 0.0130636	prefixspan 0.0130272	tremendous 0.0154127	without 0.0175662	pass 0.011357	size 0.0114699
	push 0.0103159	pseudo 0.0109016	innovate 0.0154127	fp-tree 0.0102471	disclose 0.011357	xml 0.010954

Table 2: Comparison of the content of theme "Frequent Pattern Mining" over different views

terminate the EM algorithm early when the average view distribution for view 0 (i.e., $\sum_{D\in\mathcal{D}} p(v_0|D,C)/|D|$) drops under a threshold, say 0.1. This gives us a good starting point for the EM algorithm. Then, we do this procedure again for multiple trials and select the best start point (i.e., the one with the highest likelihood). Finally, we run the EM algorithm beginning with this selected start point until it converges. The results for this experiment are selectively presented in the following table.

In Table 2, we see that the content of this selected theme varies over different views. From the global view, in which all documents are included, we can tell that this theme is talking about frequent pattern mining. From the view of Author A, we see specific frequent pattern mining techniques such as database projection, apriori, prefixspan, and closet. From the view of Author B, we see that he is not as deep into techniques of mining frequent patterns, but rather more associated with introductional and innovated work of frequent pattern mining. From the view of the years before 1993, the corresponding theme barely has any connection to frequent pattern mining. This is reasonable however, since the first and most influential paper of frequent pattern mining was published in 1993. From the view of year 1993 to 1999, we see that this theme evolutes to talk about association rules, which is perhaps the most important application of frequent pattern mining at that time. Specific techniques, such as fdm (Fast Distributed Mining of associate rules) appears high in the word distribution. From the view of the years after 1999, it is interesting to see the appearance of more new applications of frequent pattern mining, such as graphs and web. The terms corresponding to specific techniques of mining graph patterns and sequential patterns, e.g., gspan and bide, are with high probabilities in the theme word distribution. In the view corresponding to a combined context (Author A and after 1999), the top terms include "close", "top-k", and "fp-tree", which well reveal the preferences of author A in frequent pattern mining. The view specific theme for the combined context "Author B after 1999" is not well associated with the global theme again, which is consistent to the fact that Author B is not activate in frequent pattern mining any more after 2000.

This experiment shows that the CPLSA model can extract and compare the theme variations over different views effectively.

4.2 Spatiotemporal theme analysis

In this experiments, we show the effectiveness of CPLSA models on spatiotemporal analysis of themes. The context features we consider in this experiment is time stamps and location information of documents. The tasks of this specific contextual text mining problem are: (1) extract global themes from the collection, which are shared by different time and locations; (2) for each time stamp, compute the distribution of theme and locations, from which we can draw the theme distribution snapshots over locations; and (3) compare the views of themes across contexts.

It is interesting to see that the second task is not a common task of CtxTM. Let a context C be denoted as (t,l) where t and l refer to time and location, the task is to estimate $p(\kappa|D,(t,l))$ for each κ, and $P(\theta,l|t)$ for any t:

$$p(\theta,l|t)=\sum_{\kappa:(t,l)\in c(\kappa)} p(\theta|\kappa)p(\kappa|t,l)p(l|t)$$

We collect 9377 MSN Space documents with a time-bounded query submitted to Google blogsearch, with the keywords "Hurricane Katrina". In this dataset, 7118 documents provide explicit location information, and the locations of others are tagged as "unknown". We segment the time stamps into six weeks, extract and compare the common themes over different locations in United States.

Each combination of the 50 States and six week consists a unique "context", which gives us $50*6 = 300$ contexts. Since there are many contexts, it is difficult to estimate all the views precisely. Since we are only interested in the strength variations of global themes over all the contexts, it is reasonable to simplify the model by assuming that the content of the global themes does not vary over contexts. Therefore, we use the FV-CPLSA model presented in Section 3.2 to model the global themes and their coverage variations over time and locations. We further assume that $p(\kappa_C|D,C)$ is a constant that controls the impact of the context on selecting the coverage of themes. By estimating the free parameters, our goal is to compute the theme-location coverage:

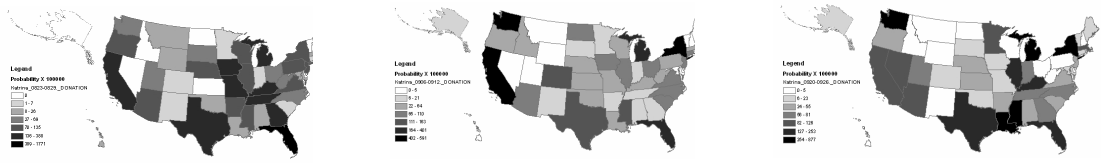
$$p(\theta,l|t)=\frac{p(\theta,l,t)}{\sum_{\theta'}\sum_{l'}p(\theta',l',t)}=\frac{p(\theta|\kappa_{t,l})p(t,l)}{\sum_{\theta'}\sum_{l'}P(\theta'|\kappa_{t,l'})P(t,l')}$$

where $p(\kappa_{t,l}|t,l)=1$, $p(t,l)$ can be computed from the word count in time period t at location l divided by the total word count in the collection.

With $p(\theta,l|t)$ computed, we can visualize the theme-location coverage by fulfill $p(\theta,l|t)$ in a snapshot map. In Figure 2, we show one of the 10 global themes we extracted from the blog dataset and its theme-location coverage at different time. From the top terms in this theme, we can infer that this theme is talking about aid and donations that were made to the hurricane affected areas. Figure 2 well demonstrates the evolution of theme-location coverage over different time periods. A detailed description of theme variation over time and location can be found in [13].

The next task is similar to the experiment in Section 4.1, which is to compare the views of themes across contexts. Specifically, we partition the states into four groups: Affected States; Peripheral States; Coast States; and Inland States. We partition the time line into spans with the length of two weeks. Then we use the FC-CPLSA model to compare the views of themes corresponding to different contexts. The results are selectively shown in Table 3.

It is easy to see that from the view of "Periphery States", the content of the theme "donation" is quite similar to the common theme extracted in Figure 2. People tend to talk about donations and supplies with food. However, from the view of "Affected Areas", which corresponds to the hur-



(a) Week1: 08/23-08/29 (b) Week Three:09/06-09/12 (c) Week Five: 09/20-09/26

Figure 2: Selected snapshots for theme "Aid and Donation" of Hurricane Katrina.

Affected States	Peripheral States	Week1-2	Week5-6
medical 0.0192	donate 0.0238	donate 0.0351	their 0.0142
comfort 0.0141	relief 0.0204	help 0.0296	help 0.0120
health 0.0137	red 0.0132	relief 0.0181	family 0.0091
ship 0.0133	cross 0.0105	red 0.0151	rebuild 0.0088
volunteer 0.0129	link 0.0086	please 0.0143	school 0.0080
hospital 0.0090	food 0.0078	cross 0.0142	children 0.0068
team 0.0081	medical 0.0074	need 0.0134	need 0.0061
assist 0.0081	supply 0.0069	volunteer 0.0120	health 0.0059
care 0.0072	charity 0.0067	victim 0.0084	evacuee 0.0057
service 0.0053	volunteer 0.0060	blood 0.0057	parish 0.0051

Table 3: Comparison of the content of the theme "Aid and Donation" over different views

ricane affected states such as Louisiana, people care more about medical aid and hospital cares. In the first two weeks, the view of this theme is still quite similar to the common theme. However in the last two weeks, we can notice that the "helps" become more about rebuilding and helping the returning evacuees.

This group of experiments show that our general model is effective to analyze spatiotemporal theme patterns.

4.3 Event Impact Analysis

In many scenarios, a collection of documents are usually associated with a series of events. For example, weblogs usually reflect people's opinions about the events happening. The research topics covered by scientific literatures are also likely to be affected by the influential related events, such as the invention of WWW, and the proposing of a new research direction. The impact of such event can usually be analyzed by comparing the themes in the documents published before versus after the event. In this experiment, we apply CPLSA on the problem of event impact analysis. Since each event gives a possible segmentation of the time line, this analysis also provides an evaluation of CPLSA on modeling overlapping views that are not orthogonal to each other. Although the experiments in previous sections also covers some overlapping views (e.g., a view corresponding to a location and a view corresponding to a time stamp), these overlaps are caused by different types of, or orthogonal context features (e.g., time and location). In reality however, the overlapping views with the same type of context feature is desirable. For example, a business analyzer may need to analyze and compare the customers' opinions in the first week, in the first month, in the first season, or in the first year after a new product is released. One strength of our model is that we allow the analysis views that overlap with each other. In this experiment, we evaluate our model on event impact analysis and overlapping view analysis.

We collect the abstracts of 1472 papers published in 28 years' SIGIR conferences from ACM Digital Library. We select two influential events to the Information Retrieval community in the 90s. One is the beginning of Text REtrieval Conferences (TREC) in 1992, which provide large-scale standard text datasets and judgements for many retrieval problems. The other is the introduction of language model into Information Retrieval in 1998, which began a genre of research and led to a lot of publications. Our goal is to use the CPLSA model to reveal the impact of these two events in IR research, i.e., how the content of research topics change after the two events.

To achieve this, we assign the abstracts in SIGIR proceedings into four contexts, each corresponds to a time span. The first context includes all the documents were published before 1993, in which is the first SIGIR conference after the start of TREC. The second context contains documents published on or after that. The third context includes abstracts before the year 1998, in which the first paper of language model in information retrieval was published. The fourth context contains all abstracts published on or after 1998. It is clear that there are overlaps between these contexts. We also include a global view, which corresponds to all the abstracts in SIGIR proceedings.

We use the same strategy as presented in Section 4.1 to avoid the EM algorithm to be trapped in unexpected local maximums. We extract 10 salient global themes from this collection and present the most interesting one.

From the global view in Table 4, we see that this theme is talking about retrieval models, especially term weighting and relevance feedback. The content of this common theme varies from different views. From the Pre-Trec view, which corresponds to the time before 1993, we see that vector space model dominates, and boolean queries are mentioned frequently. In the Post-Trec view, however, we notice that XML retrieval model has been paid more attention to. Also, we see specific types of data (email) and other terms related to the nature of TREC (e.g., collect, judgement, rank). It is more interesting when comparing the view "Pre-Language Model" and "Post-Language Model". We see that before 1998, the retrieval models are dominated by probabilistic models. After 1998, however, it is very clear that language model dominates the theme. The top ranked terms have changed to indicate language models, parameter estimations, likelihood and probability distributions, and language model smoothing. This is consistent with our prior knowledge. The overlapping views, for example Pre-LM and Pre-Trec, do share some content but clearly with different focuses. Pre-Trec, which is more faraway, emphasizes vector space model while Pre-LM emphasizes probabilistic models. This experiment shows that our method is effective to analyze event impact and model the overlapping views.

5. RELATED WORK

The most relevant work is the Probabilistic Latent Semantic Analysis model (PLSA) proposed by Hofmann [7, 8], which models a document as a mixture of aspects, where each aspect is represented by a multinomial distribution over the whole vocabulary. Our CPLSA model is a natural extension of PLSA to incorporate context. To avoid overfitting in PLSA, Blei and co-authors proposed a generative aspect model called Latent Dirichlet Allocation (LDA), which could

Views:	Global	Pre-Trec	Post-Trec	Pre-Language Model	Post-Language Model
SIGIR	term 0.159983	vector 0.0514067	xml 0.0677684	probabilist 0.0777954	model 0.16867
	relevance 0.0751814	concept 0.0297583	element 0.0212121	model 0.0431573	language 0.0752643
	weight 0.0659849	extend 0.0297405	email 0.0197383	logic 0.0403557	estimate 0.0520434
	feedback 0.0372254	model 0.0291697	collect 0.0191258	ir 0.0337741	parameter 0.0281169
	independence 0.031063	space 0.0236088	locate 0.0187425	boolean 0.028073	distribution 0.0268227
	model 0.0309212	boolean 0.0151455	judgment 0.0140086	fuzzy 0.0201544	probable 0.0205655
	frequent 0.0233021	function 0.0123171	rank 0.010205	algebra 0.0199632	smooth 0.0197662
	probabilist 0.018762	u 0.00898533	overlap 0.00975133	probable 0.0124902	score 0.0166799
	document 0.0173198	feedback 0.00860945	contextual 0.00936265	estimate 0.0119202	retrieval 0.0137085
	assume 0.0172082	specify 0.0083182	solution 0.00913	weight 0.0111257	markov 0.0118979
	dependency 0.0157547	correlate 0.00779721	subtopic 0.00791172	rank 0.0107045	likelihood 0.00585364

Table 4: Comparison of theme content over different views in SIGIR collection

also extract a set of themes from a document collection [2]. LDA, however, does not model context either. Although we have not explored it, one can also make LDA contextualized in the same way as we have done to PLSA in this paper. Recently, some extensions of this work have considered some specific types of context. For example, temporal context is considered in [6, 16, 4, 14]. Multi-collection context is analyzed in [18]. Author-topic analysis is proposed in [17]. Li et al. proposed a probabilistic model to detect retrospective news events by explaining the generation of "four Ws[1]" from each news article [11]. Our work is a generalization of these studies of specific context and provides a general probabilistic model which can be applied to all kinds of context.

Temporal context is also addressed in Kleinberg's work on discovering bursty and hierarchical structures in streams [9] and some work on topic/event/trend detection and tracking (e.g., [1, 3, 12, 10, 15]). However, most of this work assumes one document only belongs to one topic and cannot be easily generalized to analyze other contexts.

6. CONCLUSIONS

In this paper, we present a study of the general problem of contextual text mining. We formally defined the basic tasks of contextual theme analysis, and proposed a novel probabilistic mixture model to extract themes and model their content and coverage variations over different, possibly overlapping contexts. The problem definition and the proposed model are quite general and cover a family of specific contextual theme analysis problems and methods as special cases. Empirical experiments on three different datasets show that the proposed model is effective for extracting the themes and comparing the views and coverages of themes across quite different contexts.

Our work is an initial step toward a general model for contextual text mining. An important future research direction is to further study how to better estimate the proposed mixture model as discussed in Section 3.1. Another important future research direction is to create evaluation criteria and judgements so that we can quantitatively evaluate different contextual text mining approaches.

7. ACKNOWLEDGMENTS

This work was in part supported by the National Science Foundation under award numbers 0425852, 0347933, and 0428472.

8. REFERENCES

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[3] S. Boykin and A. Merlino. Machine learning of event segmentation for news on demand. *Commun. ACM*, 43(2):35–41, 2000.

[4] 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.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statist. Soc. B*, 39:1–38, 1977.

[6] T. L. Griffiths and M. Steyvers. Fiding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl.1):5228–5235, 2004.

[7] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of UAI'99*.

[8] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of ACM SIGIR'99*.

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

[10] A. Kontostathis, L. Galitsky, W. M. Pottenger, S. Roy, and D. J. Phelps. A survey of emerging trend detection in textual data mining. *Survey of Text Mining*, pages 185–224, 2003.

[11] Z. Li, B. Wang, M. Li, and W.-Y. Ma. A probabilistic model for retrospective news event detection. In *Proceedings of SIGIR'05*, pages 106–113, 2005.

[12] J. Ma and S. Perkins. Online novelty detection on temporal sequences. In *Proceedings of KDD'03*, pages 613–618, 2003.

[13] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of WWW '06*, pages 533–542, 2006.

[14] 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.

[15] R. Nallapati, A. Feng, F. Peng, and J. Allan. Event threading within news topics. In *Proceedings of CIKM'04*, pages 446–453, 2004.

[16] 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.

[17] 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.

[18] 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.

[1] who, when, where and what (keywords)