A Survey of Topic Modeling in Text Mining

Rubayyi Alghamdi
Information Systems Security
CIISE, Concordia University
Montreal, Quebec, Canada

Khalid Alfalqi
Information Systems Security
CIISE, Concordia University
Montreal, Quebec, Canada

***Abstract*—Topic models provide a convenient way to analyze large of unclassified text. A topic contains a cluster of words that frequently occur together. A topic modeling can connect words with similar meanings and distinguish between uses of words with multiple meanings. This paper provides two categories that can be under the field of topic modeling. First one discusses the area of methods of topic modeling, which has four methods that can be considerable under this category. These methods are Latent semantic analysis (LSA), Probabilistic latent semantic analysis (PLSA), Latent Dirichlet allocation (LDA), and Correlated topic model (CTM). The second category is called topic evolution models, which model topics by considering an important factor time. In the second category, different models are discussed, such as topic over time (TOT), dynamic topic models (DTM), multiscale topic tomography, dynamic topic correlation detection, detecting topic evolution in scientific literature, etc.**

Keywords*—Topic Modeling; Methods of Topic Modeling; Latent semantic analysis (LSA); Probabilistic latent semantic analysis (PLSA); Latent Dirichlet allocation (LDA); Correlated topic model (CTM); Topic Evolution Modelin

I. INTRODUCTION

To have a better way of managing the explosion of electronic document archives these days, it requires using new techniques or tools that deals with automatically organizing, searching, indexing, and browsing large collections. On the side of today's research of machine learning and statistics, it has developed new techniques for finding patterns of words in document collections using hierarchical probabilistic models. These models are called "topic models". Discovering of patterns often reflect the underlying topics that united to form the documents, such as hierarchical probabilistic models are easily generalized to other kinds of data; topic models have been used to analyze things rather than words such as images, biological data, and survey information and data [1].

The main importance of topic modeling is to discover patterns of word-use and how to connect documents that shared similar patterns. So, the idea of topic models is that term which can be working with documents and these documents are mixtures of topics, where a topic is a probability distribution over words. In other word, topic model is a generative model for documents. It specifies a simple probabilistic procedure by which documents can be generated.

Create a new document by choosing a distribution over topics. After that, each word in that document could choose a topic at random depends on the distribution. Then, draw a word from that topic. [2]

On the side of text analysis and text mining, topic models rely on the bag-of-words assumption which is ignoring the information from the ordering of words. According to Seungil and Stephen, 2010, "Each document in a given corpus is thus represented by a histogram containing the occurrence of words. The histogram is modeled by a distribution over a certain number of topics, each of which is a distribution over words in the vocabulary. By learning the distributions, a corresponding low-rank representation of the high-dimensional histogram can be obtained for each document" [3] The various kind of topic models, such as Latent semantic analysis (LSA), Probabilistic latent semantic analysis (PLSA), Latent Dirichlet allocation (LDA), Correlated topic model (CTM) have successfully improved classification accuracy in the area of discovering topic modeling [3].

When time passes by, topics in a document corpus evolve, modeling topics without considering time will confound topic discovery. Modeling topics by considering time is called topic evolution modeling. Topic evolution modeling can disclose important hidden information in the document corpus, allowing identifying topics with the appearance of time, and checking their evolution with time.

There are a lot of areas that can use topic evolution models. A typical example would be like this: a researcher wants to choose a research topic in a certain field, and would like to know how this topic has evolved over time, and try to identify those documents that explained the topic. In the second category, paper will review several important topic models.

These two categories have a good high-level view of topic modeling. In fact, they are helpful ways to better understanding the concepts of topic modeling. In addition, it will discuss inside each category. For example, the four methods that topic modeling rely on are Latent semantic analysis (LSA), Probabilistic latent semantic analysis (PLSA), Latent Dirichlet allocation (LDA), Correlated topic model (CTM). Each of these methods will have a general overview, the importance of these methods and an example that can describe the general idea of using this method. On the other hand, paper will mention the areas that topic modeling evolution provides such as topic over time (TOT), dynamic topic models (DTM), multiscale topic tomography, dynamic topic correlation detection, detecting topic evolution in scientific literature and the web of topics. Furthermore, it will going to present the overview of each category and provides examples if any and some limitations and characteristics of each part.

This paper is organized as follows. Section II provides the first category methods of topic modeling with its four methods and their general concepts as subtitles. Section III overviews of second category which is topic modeling evolution including its parts. Then it is followed by conclusions in Section IV.

II. THE METHODS OF TOPIC MODELING

In this section, paper will discuss some of the topic modeling methods that deals with words, documents and topics. In addition, the general idea of each of these methods, and present some example for these methods if any. Also, these methods involve in many applications so it will have a brief idea in what applications that can these methods work with.

A. Latent Semantic Analysis

Latent semantic analysis (LSA) is a method or a technique in the area of Natural language processing (NLP). The main goal of Latent semantic analysis (LSA) is to create vector based representation for texts' to make semantic content. By vector representation (LSA) computes the similarity between texts' to pick the heist efficient related words. In the past LSA was named as latent semantic indexing (LSI) but improved for information retrieval tasking. So, finding few documents that close to the query that given from many documents. LSA should have many aspects to give approach such as key words matching, Wight key words matching and vector representation depends on occurrences of words in documents. Also, Latent semantic analysis (LSA) uses singular value decomposition (SVD) to rearrange the data.

SVD is a method that uses a matrix to reconfigure and calculates all the diminutions of vector space. In addition, the damnations In vector space will be computed and organized from most to the least Important .in LSA the most significant assumption will be used to fined the meaning of the text otherwise least important will be ignored in the assumption .By searching about words that have a high rate of similarity will be occurred if that wards have similar vector. To describe the most essential steps in LSA first collecting a huge set of relevant text then divide it by documents. Second make co occurrence matrix for terms and documents also giving the cell name such as documents x, terms y and m for dimensional value for terms and n dimensional vector for documents. Third each cell will be whetted and calculated finials SVD will play big roil to compute all the diminutions and make three matrices.

B. Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (PLSA) is an approach that has been release after LSA method to fix some disadvantages that have found into LSA. Jan Puzicha and Thomas Hofmann introduced it in 1999. PLSA is a method that can be automated document indexing which is based on a statistical latent class model for factor analysis of count data, and also this method tries to improve the Latent Semantic Analysis (LSA) in a probabilistic sense by using a generative model. The main goal of PLSA is that identifying and distinguishing between different contexts of word usage without recourse to a dictionary or thesaurus. It includes two important implications: First one, it allows to disambiguate polysemy, i.e., words with multiple meanings. Second thong, it discloses topical similarities by grouping together words that shared a common context [3].

According to Kakkonen, Myller, Sutinen, and Timonen, 2008, "PLSA is based on a statistical model that is referred as *an aspect model*. An *aspect model* is a latent variable model for co-occurrence data, which associates unobserved class variables with each observation" [4]. The PLSA method comes to improve the method of LSA, and also to solve some other problems that LSA cannot solve it. PLSA has been successful in many real-world applications, including computer vision, and recommender systems. However, since the number of parameters grows linearly with the number of documents, PLSA suffers from overfitting problems. Even though, it will discuss some of these applications later [5].

In the other hand, PLSA based on algorithm and different aspects. In this probabilistic model, it introduce a Latent variable zk ∈ {z1, z2,..., zK}, which corresponds to a potential semantic layer. Thus, the full model: p (di) on behalf of the document in the data set the probability; p (wj | zk) zk representatives as defined semantics, the related term (word) of the opportunities are many; p (zk | di) represents a semantic document distribution. Using these definitions, will generate model use it to generate new data by the following steps: [3]

1) *Select a document di with probability P (di),*
2) *Pick a latent class zk with probability P (zk | di),*
3) *Generate a word wj with probability P (wj |zk).*

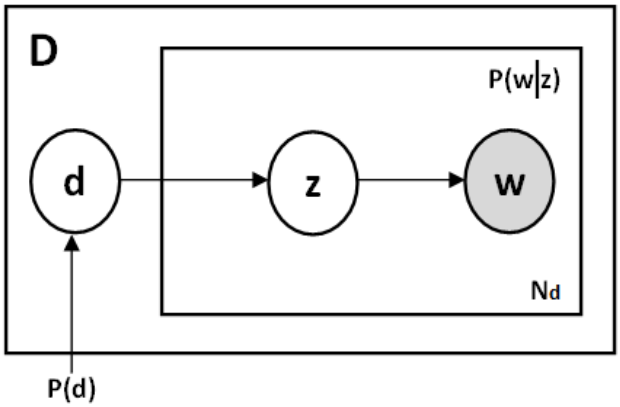

Fig. 1. High-Level View of PLSA

PLSA has two different formulations that can present this method. The first formulation is symmetric formulation, which will help to get the word (*w)* and the document (*d)* from the latent class *c* in similar ways. By using the conditional probabilities *P*(*d* | *c*) and *P*(*w* | *c*). The second formulation is the asymmetric formulation. In this formulation each document *d*, a latent class is chosen conditionally to the document according to *P*(*c* | *d*), and the word can be generated from that class according to *P*(*w* | *c*) [6]. Each of these two formulations has rules and algorithms that could be used for different purposes. These two formulations have been improved right now and this was happened when they released the Recursive Probabilistic Latent Semantic Analysis

(RPLAS). This method is extension for the PLAS; also it was improving for the asymmetric and symmetric formulations.

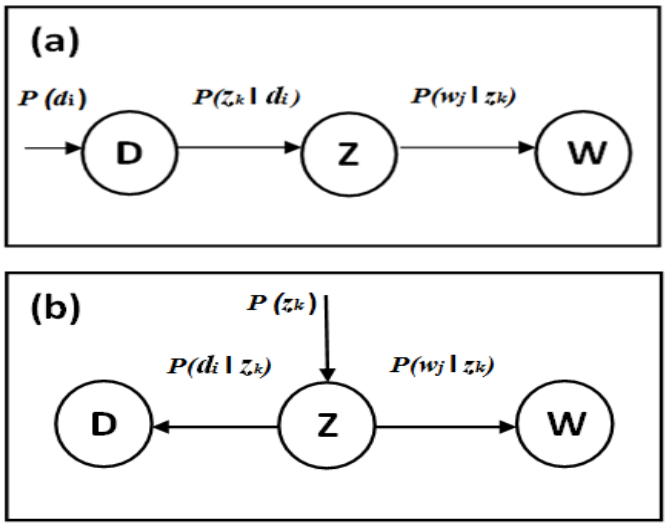

Fig. 2. A graphical model representation of the aspect model in the asymmetric (a) and symmetric (b) parameterization [3]

In the term of PLSA applications, PLSA has applications in many various field such as information retrieval and filtering, natural language processing and machine learning from text. In specific, some of these applications are automatic essay grading, classification, topic tracking, image retrieval and automatic question recommendation. Will discuss two of these applications as follows:

- Image retrieval: PLSA model has the visual features that it uses to represent each image as a collection of visual words from a discrete and finite visual vocabulary. Having an occurrence of visual words in an image is hereby counted into a co-occurrence vector. Each image has the co-occurrence vectors that can help to build the co-occurrence table that is used to train the PLSA model. After knowing the PLSA model, can apply the model to all the images in the database. Then, the pediment of the vector is to represent it for each image, where the vector elements denote the degree to which an image depicts a certain topic [7].
- Automatic question recommendation: One of the significant application that PLSA deal with is question recommendation tasks, in this kind of application the word is independent of the user if the user want a specific meaning, so when the user get the answers and the latent semantics under the questions, then he can make recommendation based on similarities on these latent semantics. Wang, Wu and Cheng, 2008 reported that "Therefore, PLSA could be used to model the users' profile (represented by the questions that the user asks or answers) and the questions as well through estimating the probabilities of the latent topics behind the words. Because the user's profile is represented by all the questions that he/she asks or answers, we only need to consider how to model the question properly" [8]

C. Latent Dirichlet Allocation

The reason of appearance of latent Dirichlet allocation (LDA) model is to improve the way of mixture models that capture the exchangeability of both words and documents from the old way by PLSA and LSA. This was happening In 1990, so the classic representation theorem lays down that any collection of exchangeable random variables has a representation as a mixture distribution—in general an infinite mixture. [9].

There are huge numbers of electronic document collections such as the web, scientific interesting, blogs and news articles literature in the recent past has posed several new, challenges to researchers in the data mining community. Especially there is growing need for automatic techniques to visualize, analyze and summarize mine these document collections. In the recent past, latent topic modeling has become very popular as a completely unsupervised technique for topic discovery in large document collections. This model, such as LDA [10]

Latent Dirichlet Allocation (LDA) is an Algorithm for text mining that is based on statistical (Bayesian) topic models and it is very widely used. LDA is a generative model that means that it tries mimics what the writing process is. So it tries to generate a document given the topic. It can also be applied to other types of data. There are tens of LDA based models including: temporal text mining, author- topic analysis, supervised topic models, latent Dirichlet co-clustering and LDA based bio-informatics [11].[18]

In the simple basic idea of the process each document is modeled as a mixture of topics, and each topic is a discrete probability distribution that defines how likely each word is to appear in a given topic. These topic probabilities give a concise representation of a document. Here, a "document" is a "bag of words" with no structure beyond the topic and word statistics.

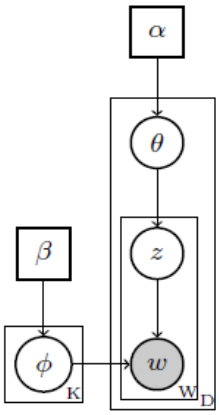

Fig. 3. A graphical model representation of LDA

LDA models each of D documents as a mixture over K latent topics, each of which describes a multinomial distribution over a W word vocabulary. Figure 3 shows the graphical model representation of the LDA model. The generative process for the basic LDA is as follows:

For each of Nj words in document j

1) *Choose a topic zij ∼ Mult(θj)*
2) *Choose a word xij ∼ Mult(φzij)*

Where the parameters of the multinomials for topics in a document θj and words in a topic φk have Dirichlet priors [12]

Indeed, there are several of applications and model based on the Latent Dirichlet Allocation (LDA) method such as:

- Role discovery: Social network analysis (SNA) is the study of mathematical models for interactions among people, organizations and groups. Because of the emergence connections among the 9/11 hijackers and the huge data sets of human on the popular web service like facebook.com and MySpace.com, there has been growing interest in social network analysis. That leads to exist of Author-Recipient-Topic (ART) model for social network analysis. The model combines the Latent Dirichlet Allocation (LDA) and the Author-Topic (AT) model. The Idea of (ART) is learn topic distributions based on the direction-sensitive messages sent between the senders and receivers [13].
- Emotion topic: The Pairwise-Link-LDA model, which is focus on the problem joint modeling of text and citations in the topic modeling area. It's builds on the ideas of LDA and Mixed Membership Stochastic Block Models (MMSB) and allows modeling arbitrary link structure [14].
- Automatic essay grading: The Automatic essay grading problem is closely related to automatic text categorization, which has been researched since the 1960s. Comparison of Dimension Reduction Methods for Automated Essay Grading. LDA has been shown to be reliable methods to solve information retrieval tasks from information filtering and classification to document retrieval and classification [15].
- Anti-Phishing: Phishing emails are ways to theater the sensitive information such as account information, credit card, and social security numbers. Email Filtering or web site filtering is not the effective way to prevent the Phishing emails. Because latent topic models are clusters of words that appear together in email, user can expect that in a phishing email the words "click" and "account" often appear together. Usual latent topic models do not take into account different classes of documents, e.g. phishing or non-phishing. For that reason the researchers developed a new statistical model, the latent Class-Topic Model (CLTOM), which is an extension of latent Dirichlet allocation (LDA) [16].
- Example of LDA: This section is to provide an illustrative example of the use of an LDA model on real data. By using the subset of the TREC AP corpus containing 16,000 documents. First, remove the stop-words in TREC AP corpus before running topic modeling. After that, use the EM algorithm to find the Dirichlet and conditional multinomial parameters for a 100-topic LDA model. The top words from some of the resulting multinomial distributions are illustrated in Figure 4. As a result, these distributions seem to capture some of the underlying topics in the corpus (it is named according to these topics [9].

"ARTS"	"BUDGET"	"CHILDREN"	"EDUCATION"
New	Million	Children	School
Film	Program	Women	Students
Show	Tax	People	Schools
Music	Budget	Child	Education
Movie	Billion	Years	Teachers
Play	Federal	Families	High
Musical	Year	Work	Public
Best	Spending	Parent	Teacher
Actor	New	Says	Bennett
First	State	Family	Manigat
York	Plan	Welfare	Namphy
Opera	Money	Men	State
Theater	Programs	Percent	President
Actress	Government	Care	Elementary
Love	Congress	Life	Haiti

Fig. 4. Most likely words from 4 topics in LDA from the AP corpus: the topic titles in quotes are not part of the algorithm

D. Correlated topic model

Correlated Topic Model (CTM) is a kind of statistical model used in natural language processing and machine learning. Correlated Topic Model (CTM) used to discover the topics that shown in a group of documents. The key for CTM is the logistic normal distribution. Correlated Topic Models (CTM) is depending on LDA.

TABLE I. THE CHARACTERISTICS OF TOPIC MODELING METHODS [17]

Name of The Methods	Characteristics
Latent Semantic Analysis (LSA)	* LSA can get from the topic if there are any synonym words. * Not robust statistical background.
Probabilistic Latent Semantic Analysis (PLSA)	* It can generate each word from a single topic; even though various words in one document may be generated from different topics. * PLSA handles polysemy.
Latent Dirichelet Allocation (LDA)	* Need to manually remove stop-words. * It is found that the LDA cannot make the representation of relationships among topics.
Correlated Topic Model (CTM)	* Using of logistic normal distribution to create relations among topics. * Allows the occurrences of words in other topics and topic graphs.

TABLE II. THE LIMITATIONS OF TOPIC MODELING METHODS [17]

Name of The Methods	Limitations
Latent Semantic Analysis (LSA)	- It is hard to obtain and to determine the number of topics. - To interpret loading values with probability meaning, it is hard to operate it.
Probabilistic Latent Semantic Analysis (PLSA)	- At the level of documents, PLSA cannot do probabilistic model.
Latent Dirichelet Allocation (LDA)	- It becomes unable to model relations among topics that can be solved in CTM method.
Correlated Topic Model (CTM)	- Requires lots of calculation - Having lots of general words inside the topics.

III. METHODS ABOUT TOPIC EVOLUTION MODELS

A. Overview of topic evolution models

When time goes by, the themes of a document corpus evolve. Modeling topics without considering time will cause problems. For example, in analyzing topics of U.S. Presidential State-of-the-Union addresses, LDA did not correctly do it by confounding Mexican-American War with some aspects of World War I, since LDA did not notice that there was 70-year separation between the two events.

It is important to model topic evolution, so people can identify topics within the context (i.e. time) and see how topics evolve over time. There are a lot of applications where topic evolution models can be applied. For example, by checking topic evolution in scientific literature, it can see the topic lineage, and how research on one topic influences on another.

This section will review several important papers that model topic evolutions. These papers model topic evolution by using different models, but all of them consider the important factor time when model topics. For example, probabilistic time series models are used to handle the issue in paper "dynamic topic models", and non-homogeneous Poisson processes and multiscale analysis with Haar wavelets are employed in paper "multiscale topic tomography" to model topic evolution.

B. A Non-Markov Continuous-Time Method

Since most of the large data sets have dynamic co-occurrence patterns, and word and topic co-occurrence patterns change over time, TOT models topics and their changes over time by taking into account both the word co-occurrence pattern and time [19]. In this method, a topic is considered as being associated with a continuous distribution over time.

In TOT, for each document, multinomial distribution over topics is sampled from Dirichlet, words are generated from multinomial of each topic, and Beta distribution of each topic generates the document's time stamp. If there exists a pattern of a strong word co-occurrence for a short time, TOT will create a narrow-time-distribution topic. If a pattern of a strong word co-occurrence exists for a while, it will generate a broad-time-distribution topic.

The main point of this paper is that it models topic evolution without discretizing time or making Markov assumptions that the state at time t + t1 is independent of the state at time t. By using this method on U.S. Presidential State-of-the-Union address of two centuries, TOT discovers topics of time-localization, and also improves the word-clarity over LDA. Another experimental result on the 17-year NIPS conference demonstrates clear topical trends.

C. Dynamic Topic Models (DTM)

The authors in this paper developed a statistical model of topic evolution, and developed approximate posterior inference techniques to decide the evolving topics from a sequential document collection [20]. It assumes that corpus of documents is organized based on time slices, and the documents of each time slice are modeled with K-component model, and topics associated with time slice *t* evolve from topics corresponding to slice time *t-1*.

Dynamic topic models estimate topic distribution at different epochs. It uses Gaussian prior for the topic parameters instead of Dirichlet prior, and can capture the topic evolution over time slices. By using this model, what words are different from the previous epochs can be predicted.

D. Multiscale Topic Tomography

This method assumes that the document collection is sorted in the ascending order, and that the document collection is grouped into equal-sized chunks, each of which represents the documents of one epoch. Each document in an epoch is represented by a word-count vector, and each epoch is associated with its word generation Poisson parameters, each of which represents the expected word counts from a topic. Non-homogeneous Poisson process was used to model word counts, since it is a natural way to do the task, and also because it is amendable to sequence modeling through Bayesian multi-scale analysis. Multi-scale analysis was also employed to the Poisson parameters, which can model the temporal evolution of topics at different time-scales.

This method is similar to DTM, but provides more flexibility by allowing studying the topic evolution with various time-scales [21].

E. A Non-parametric Approach to Dynamic Topic Correlation Detection (DCTM)

This method models topic evolution by discretizing time [22]. In this method, each corpus contains a set of documents, each of which contains documents with the same timestamp. It assumes that all documents in a corpus share the same time-scale, and that each document corpus shares the same vocabulary of size d.

Basically, DCTM maps the high-dimensional space (words) to lower-dimensional space (topics), and models the dynamic topic evolution in a corpus. A hierarchy over the correlation latent space is constructed, which is called temporal prior. The temporal prior is used to capture the dynamics of topics and correlations.

DCTM works as follows: First of all, for each document corpus, the latent topics are discovered, and this is done by first summarizing the contribution of documents at certain

time, which is done by aggregating the features in all documents. Then, Gaussian process latent variable model (GP-LVM) is used to capture the relationship between each pair of document and topic set. Next, hierarchical Gaussian process latent variable model (HGP-LVM) is employed to model the relationship between each pair of topic sets. They also use the posterior inference of topic and correlations to identify the dynamic changes of topic-related word probabilities, and to predict topic evolution and topic correlations.

An important feature of this paper is that it is non-parametric model since it can marginalize out the parameters, and it exhibits faster convergence than the generative processes.

F. Detecting Topic Evolution of Scientific Literature

This method employs the observation that citation indicates important relationship between topics, and it uses citation to model topic evolution of scientific literature [23]. Not only papers that are in a corpus D(t) are considered for topic detection, but those papers that are cited are also taken into account. It uses Bayesian model to identify topic evolution.

In this method, "a document consists of a vocabulary distribution, a citation and a timestamp". Documents corpus is divided into a set of subsets based on the timestamp, for time unit *t*, the corresponding documents are represented with D(t). For each time unit, topics are generated independently. The topic evolution analysis in this paper is specified to analyze the relationship between topics in D(t) and those in D(t-1). In other words, it models topic evolution by discretizing time.

They first proposed two citation-unaware topic evolution learning methods for topic evolution: independent topic evolution learning method and accumulative topic evolution learning method. In independent topic evolution learning method, topics in D(t) are independent from those in D(t-1), while in accumulative topic evolution learning method, topics in D(t) are dependent on those in D(t-1). Then, Citation is integrated into the above two approaches, which is an iterative learning process based on Dirichlet prior smoothing. The iterative learning process takes into account the fact that different citations have different importance on topic evolution. Finally an inheritance topic model is proposed to capture how citations can be employed to analyze topic evolution.

G. Discovering the Topology of Topics

A topic is semantically coherent content that is shared by a document corpus. When time passes, some documents in a topic may initiate a content that is differ obviously from the original content. If the initiated content is shared by a lot of later documents, the content is identified as a new topic. This paper is to discover this evolutionary process of topics. In this paper, a topic is defined as "a quantized unit of evolutionary change in content".

This method develops an iterative topic evolution learning framework by integrating Latent Dirichlet Allocation into citation network. It also develops an inheritance topic model by using citation counts.

It works as follows: first, it tries to identify a new topic by identifying significant content changes in a document corpus. If the new content is obviously different from the original content, and the new content is shared by later documents, the new content is identified as a new topic.

The next step is to explore the relationship between the new topics and the original topics. It works by finding member documents of each topic, and examining the relationship. It also uses citation relationship to find member documents of each topic. That is, if a paper cites the start paper, this paper will be the member paper of the start paper. In addition, those papers that are textually close to the start paper; they are also member papers of the start paper. The relationship between the original topics and the new discovered topics is identified by using citation count in this paper. Their experimental results demonstrate that citations can better understand topic evolutions.

H. Summary of topic evolution models

This paper summarizes the main characteristics of topic evolution models discussed in section 3, which is listed as follows:

TABLE III. THE MAIN CHARACTERISTICS OF TOPIC EVOLUTION MODELS

Main characteristics of models	Models
Modeling topic evolution by continuous-time model	1)"Topics over time: a non-markov continuous-time model of topical trends"
Modeling topic evolution by discretizing time	1)"Dynamic topic models" 2)"Multiscale topic tomography" 3)*"*ANon-parametric Approach to Pair-wise Dynamic Topic Correlation Detection*"*
Modeling topic evolution by using citation relationship as well as discretizing time	1) "Detecting topic evolution in scientific literature: How can citations help" 2) "The Web of Topics: Discovering the Topology of Topic Evolution in a Corpus"

I. Comparison of Two Categories

The main difference of the two categories that model topics is that the first category model topics without considering time and the methods in the first category mainly model words.

The methods in the second category model topics considering time, by using continuous-time, by discretizing time, or by combining time discretization and citation relationship. Due to the different characteristics of these two categories, the methods in the second category are more accurate in terms of topic discovery.

IV. CONCLUSION

This survey paper, presented two categories that can be under the term of topic modeling in text mining. In the first

category, it has discussed the general idea about the four topic modeling methods including Latent semantic analysis (LSA), Probabilistic latent semantic analysis (PLSA), Latent Dirichlet allocation (LDA), and Correlated topic model (CTM). In addition, it explained the different between these four methods in term of characteristics, limitations and the theoretical backgrounds. Paper does not go into specific details of each of these methods. It only describes the high-level view of these topics that relate it to topic modeling in text mining. Furthermore, it has mentioned some of the applications that have been involved in these four methods. Also, it has mentioned that each method of these four is improving for the old one and modifies some of disadvantages that have found in the previous methods. Model topics without taking into account time will confound the topic discovery. In the second category, paper has discussed the topic evolution models, which model topics by considering time. Several papers use different methods to model topic evolution. Some of them model topic evolution by discretizing time, some of them use continuous-time model, and some of them employ citation relationship as well as time discretization to model topic evolution. All of these papers model topics by considering the important factor time.

REFERENCES

[1] Blei, D.M., and Lafferty, J. D. "Dynamic Topic Models", *Proceedings of the* 23rd *International Conference on Machine Learning*, Pittsburgh, PA, 2006.

[2] Steyvers, M., and Griffiths, T. (2007). "Probabilistic topic models". In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum

[3] Hofmann, T., "Unsupervised learning by probabilistic latent semantic analysis", *Machine Learning*, 42 (1), 2001, 177- 196.

[4] Kakkonen, T., Myller, N., Sutinen, E., and Timonen, J., "Comparison of Dimension Reduction Methods for Automated Essay Grading", Educational Technology &Society, 11 (3), 2008, 275-288.

[5] Liu, S., Xia, C., and Jiang, X., "Efficient Probabilistic Latent Semantic Analysis with Sparsity Control", *IEEE International Conference on Data Mining*, 2010, 905-910.

[6] Bassiou, N., and Kotropoulos C. "RPLSA: A novel updating scheme for Probabilistic Latent Semantic Analysis", *Department of Informatics, Aristotle University of Thessaloniki, Box 451 Thessaloniki 541 24, Greece* Received 14 April 2010.

[7] Romberg, S., Hörster, E., and Lienhart, R., "Multimodal pLSA on visual features and tags", *The Institute of Electrical and Electronics Engineers Inc.*, 2009, 414-417.

[8] Wu, H., Wang, Y., and Cheng, X., "Incremental probabilistic latent semantic analysis for automatic question recommendation", *ACM New York, NY, USA*, 2008, 99-106.

[9] Blei, D.M., Ng, A.Y., and Jordan, M.I., "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, 3, 2003, 993-1022.

[10] Ahmed,A., Xing,E.P., and William W. "Joint Latent Topic Models for Text and Citations", ACM New York, NY, USA, 2008.

[11] Zhi-Yong Shen,Z.Y., Sun,J., and Yi-Dong Shen,Y.D., "Collective Latent Dirichlet Allocation", Eighth IEEE International Conference on Data Mining, pages 1019–1025, 2008.

[12] Porteous, L.,Newman,D., Ihler, A., Asuncion, A., Smyth, P., and Welling, M., "Fast Collapsed Gibbs Sampling For Latent Dirichlet Allocation", ACM New York, NY, USA, 2008.

[13] McCallum, A., Wang, X., and Corrada-Emmanuel, A., "Topic and role discovery in social networks with experiments on enron and academic email", *Journal of Artificial Intelligence Research*, 30 (1), 2007, 249-272.

[14] Bao, S., Xu, S., Zhang, L., Yan, R., Su, Z., Han, D., and Yu, Y., "Joint Emotion-Topic Modeling for Social Affective Text Mining", Data Mining, 2009. ICDM '09. Ninth IEEE International Conference, 2009, 699-704.

[15] Kakkonen, T., Myller, N., and Sutinen, E., "Applying latent Dirichlet allocation to automatic essay grading", *Lecture Notes in Computer Science*, 4139, 2006, 110-120.

[16] Bergholz, A., Chang, J., Paaß, G., Reichartz, F., and Strobel, S., "Improved phishing detection using model-based features", 2008.

[17] Lee,S., Baker,J., Song,J., and Wetherbe, J.C., "An Empirical Comparison of Four Text Mining Methods", Proceedings of the 43rd Hawaii International Conference on System Sciences, 2010.

[18] X. Wang and A. McCallum. "Topics over time: a non-markov continuous-time model of topical trends". In International conference on Knowledge discovery and data mining, pages 424–433, 2006.

[19] D. M. Blei and J. D. Lafferty. Dynamic topic models. In International conference on Machine learning, pages 113–120, 2006.

[20] R. M. Nallapati, S. Ditmore, J. D. Lafferty, and K. Ung. Multiscale topic tomography. In Proceedings of KDD'07, pages 520–529, 2007.

[21] *A Non-parametric Approach to Pair-wise Dynamic Topic Correlation Detection*. In Proceedings of the Eighth IEEE International Conference on Data Mining (ICDM 2008), Pisa, Italy. December 2008.

[22] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and C. L. Giles. Detecting topic evolution in scientific literature: How can citations help? In CIKM, 2009.

[23] Yookyung Jo, John E. Hopcroft, and Carl Lagoze. The Web of Topics: Discovering the Topology of Topic Evolution in a Corpus, The 20th International World Wide Web Conference, 2011.