AN EMPIRICAL COMPARISON OF FOUR TEXT MINING METHODS*

SANGNO LEE
Texas Tech University
Lubbock, Texas 79409

JAEKI SONG
Texas Tech University, Lubbock, TX 79409
Sogang University, Seoul, Korea

YONGJIN KIM
Sogang University
Seoul, Korea

ABSTRACT

The amount of textual data that is available for researchers and businesses to analyze is increasing at a dramatic rate. This reality has led IS researchers to investigate various text mining techniques. This essay examines four text mining methods that are frequently used in order to identify their characteristics and limitations. The four methods that we examine are (1) latent semantic analysis, (2) probabilistic latent semantic analysis, (3) latent Dirichlet allocation, and (4) correlated topic model. We review these four methods and compare them with topic detection and spam filtering to reveal their peculiarity. Our paper sheds light on the theory that underlies text mining methods and provides guidance for researchers who seek to apply these methods.

Keywords: text mining, vector space model; latent semantic analysis, probabilistic latent semantic analysis; latent Dirichlet allocation; correlated topic model.

1. INTRODUCTION

Modern information systems allow firms to capture vast amounts of data. Much of this data is structured data that can be analyzed using traditional database software. Increasingly, however, large amounts of data such as textual data are unstructured, and defy simple attempts to make sense of it. Manual analysis of this unstructured textual data is increasingly impractical, and as a result, text mining methods are being developed to automate the process of analyzing this data. Text mining has been used to identify intellectual core in the information systems discipline [35, 47], topic discovery [40], customer relationship management [15], and target advertising [55]. Elsewhere text mining has been used not only to identify sentiments of investors such as negative or positive opinions from business news papers and financial websites [17, 51], but also to detect objects from images [48]. Given the need to analyze and understand textual data, and given the increasing popularity of text mining methods, the time seems right for a research essay to examine the major methods available for text mining, and to address the topic of how to appropriately apply text mining techniques in IS research and practice.

This paper has two intended contributions. First, we review four methods that are commonly used in text mining research in order to open up the use of these methods to a wider audience of researchers. The four methods that we discuss are latent semantic analysis, probabilistic latent semantic analysis, latent Dirichlet allocation, and correlated topic model. These methods range from early approaches of text mining that use discriminative models to more recent approaches that use generative or probabilistic models. We choose the four methods because of their wide applications and primary methods in the development of text mining, but do not consider syntactical and morphological approach of natural language because of the lagged development and relatively unsuccessful performance [43]. We explain that the more recently-proposed generative models overcome the limitations of previous methods and provide researchers with the ability to identify patterns in textual data through the text modeling process. Also, we present the appropriate condition in which each method could be applied by reviewing theoretical backgrounds of four methods.

Second, we present two examples for the empirical comparison of four text mining methods in the areas of topic detection and spam filtering. We examine the characteristics and limitations of each method with a descriptive way in the topic detection and with a objective way in the spam filtering. Each of the four text mining methods that we present has their own unique features, and the effective utilization of features plays a crucial role to get high performance. Thus, our comparison provides guidance for researchers and practitioners who apply text mining methods.

The paper is organized as follows. In section two, we provide an overview of fundamental concepts and assumptions in text mining. We also explain that the vector space model can be used as a basic measure of language distance. In section three, we explain the four various methods of text mining and describe the theoretical underpinnings of each method. We also explain the characteristics and limitations of each of the four methods. In section four, we illustrate the use of these four methods with topic detection and spam filtering. Finally, in section five, we provide directions for future research and review our contributions.

2. FUNDAMENTAL CONCEPTS IN TEXT MINING

The four methods of text mining that we describe here are built upon a set of basic assumptions and also upon the vector space model (VSM). VSM represents textual data in vector space and measures the distance of language vectors [44].

2.1. Concepts and assumptions

The fundamental unit of text is a *word*. Words are comprised of characters, and are the basic units from which meaning is constructed. By combining a *word* with grammatical structure,

* *This study was supported by a grant from Sogang Business School's World Class University Project (R31-20002) funded by the Korea Research Foundation.*

Received: November 16, 2009 Revised: February 17, 2010 Accepted: March 1, 2010

a *sentence* is made. Sentences are the basic unit of action in text, containing information about the action of some subject. *Paragraphs* are the fundamental unit of composition [49] and contain a related series of ideas or actions. As the length of text increases, additional structural forms become relevant, often including sections, chapters, entire documents, and finally, a *corpus* of documents. In text mining study, a *document* is generally used as the basic unit of analysis because a single writer commonly writes the entirety of a document and the document discusses a single topic. A document can be a paper, an essay, or a book, depending on the type of analysis being performed and depending upon the goals of the researcher. In some cases, a document may contain only a chapter, a single paragraph, or even a single sentence.

In text mining studies, the syntactical structure of a sentence or paragraph is intentionally ignored in order to efficiently handle the text. As a result, a sentence is regarded simply as a set of words, or a "*bag of words*" in the parlance of text mining. This idea that a sentence is a bag of words without additional structure is a basic assumption of most text mining. Because the syntactical structure of text is ignored in text mining, the order of words can be changed without impacting the outcome of the analysis. The bag-or-words concept is also refereed to as *exchangeability* in the generative language model [2].

Two closely-related problems that arise in analyzing text are *synonymy* and *polysemy.* Synonymy denotes different words but the meaning of them is identical or very similar. Polysemy is a word with multiple meanings. This problem has been addressed in a variety of ways [28]. Because different text mining methods have slightly different approaches to deal with synonymy and polysemy, we will return to this issue as we explain each of the text mining methods.

In the next section, we will review vector space model, another fundamental concept that underlies in many text mining approaches.

2.2. Vector space model

The vector space model (VSM) regards a document as a bag of words, and is mainly used for information retrieval tasks, such as keyword searches. In this model, words in documents are represented with mathematical vectors, which are one-dimensional arrays. Alternatively, the importance of a word in documents is evaluated with word frequency (*tf*) and inverse document frequency (*idf*). This scheme (*tf-idf*) is designed to consider the discriminative power of a word both *within* a document and *over* documents. The degree of similarity between two documents is calculated with cosine function which is a familiar mathematical function from Euclidian geometry, cosine $\theta = (v_1 v_2) / (||v_1|| \; ||v_2||)$. Thus, VSM represents documents with a simple and manageable matrix form, and allows distance measurement between documents. These two features are foundational for text mining methods that we will discuss in the next section.

VSM has four primary limitations. First, in information retrieval, a long lengthy document gets a low similarity to a query because the normalized value of the document, $||D||$, becomes high. As a result, a long-length document has little opportunity to match a query. Second, as we have noted earlier, the order of words in a document is still ignored because of the bag of words assumption. The syntactic structure of a document is potentially valuable information. Third, keywords in a query have to be exactly matched with words in documents, and thus the issue of synonymy is not addressed. Fourth and finally, the issue of polysemy is not addressed because VSM only considers word form. To overcome the synonyms issue, latent semantic analysis has been developed. We will turn our attention to this method.

3. FOUR METHODS IN TEXT MINING

Text mining has been developed from discriminative models such as latent semantic analysis to generative models such as probabilistic latent semantic analysis, latent Dirichlet allocation, and correlated topic model. In the next section, we first examine a discriminative model, and then other three generative models. After that, we discuss the theoretical understanding for performance difference of four methods.

3.1. Latent semantic analysis

As we have noted, VSM cannot deal with synonymy and polesemy with the following three reasons. To address these issues, latent semantic analysis (LSA) has been developed [18, 34, 39]. LSA projects an original vector space or term-document matrix into a small *factor* space. The dimensional reduction of a matrix is accomplished using singular value decomposition which decomposes an original matrix into three matrixes, a document eigenvector matrix, an eigenvalue matrix, and a term eigenvector matrix. In turn, an original matrix can be approximated by multiplying these three matrixes with only high eigenvalues. Also, topics or factors can be extracted using factor loadings and matrix rotation [34]. Because of orthogonal characteristic of factors, words in a factor have little relations with words in other factors, but words in a factor have high relations with words in that factor.

The factors in LSA can deal with more synonymy than polesemy. Words in a factor can be regarded as synonyms because words within a factor have high correlation each other. However, for polysemy, the same form of a word has to be appeared in different topics according to usage of the word. For instance, a *java* word is needed to be shown up not only in a programming language factor but also in a coffee factor. Unfortunately, the orthogonal characteristic of LSA prevents multiple occurrences of a word from different topics.

Papadimitriou et al. [39] investigate appropriate conditions for applying LSA and find the following three appropriate conditions. First, documents have the same writing style. For instance, in the official documents, *automobile* or *vehicle* is generally used instead of *car*. Second, each document is only on a single topic. Third, a word has high probability in one topic, but low probability in other topics. LSA is applied in many different areas such as information retrieval, topic detection, and essay grading, and becomes a baseline of performance of advanced methods. The second row in Table 1 shows its applications and performance.

LSA improves VSM by considering synonyms but has three limitations. First, it is a straightforward dimension reduction of a matrix and is not built upon robust probability theory. Second, the adequate number of factors cannot be determined statistically, and the determination of factor numbers depends upon human judgment. Third, polysemy is partially dealt with in LSA because of the orthogonal characteristic. To overcome the straightforward dimensional reduction of a matrix and the polesemy issue, the following generative methods have been developed, and in the next section, we explorer the first generative model, probabilistic latent semantic analysis.

3.2. Probabilistic latent semantic analysis

Under the assumption of *exchangeability*, the occurrence of words can be modeled using probabilistic theory. Unlike discriminative models, generative methods set *a model* firstly, and then estimate unknown parameters using a word-document matrix. Thus, the major differences among generative models arise from modeling methods for documents and estimation methods for parameters.

The probabilistic latent semantic analysis (PLSA) assumes that documents are generated throughout the following three steps. First, a document d is generated or selected with probability $P(d)$. Second, topic z is picked with probability $P(z|d)$. Third, each word w in a topic is generated with probability $P(w|z)$. Then,

TABLE 1. The applications of four text mining methods and their performance

Methods	**Applications**	**Performance & comments**
LSA	Search & retrieval	
	• (Herdiyeni 2009) [24]	Automatic image annotation.
	• (Chen et al. 2008) [14]	Medical image retrieval
	Automatic essay grading	
	• (Kakkonen et al. 2006) [29]	LSA > PLSA
	• (Kakkonen et al. 2008) [30]	LSA > PLSA, LDA
	Spam filtering	
	• (Gansterer et al. 2008) [21]	VSM > LSA
	• (Sun et al. 2008) [50]	LSA + SHA > KNN[1)]
	Topic detection	
	• (Sidorova et al. 2008) [47]	Abstracts data of journals
PLSA	Automatic essay grading	
	• (Kakkonen 2006) [29]	LSA > PLSA, PLSA-C
	Classification	
	• (Ahrendt et al. 2005) [1]	Music Genre
	• (Bosch, 2006) [11]	Scene classification (PLSA + KNN)
	Topic tracking	
	• (Chou & Chen 2008) [16]	Online event analysis (IPLSI)
	• (Molgaard & Larsen, 2009) [38]	
	Image retrieval	
	• (Romberg et al. 2009) [42]	High level of visual features
	Automatic question recommendation	
	• (Wu et al. 2008) [53]	Incremental PLSA
LDA	Automatic essay grading	
	• (Kakkonen, 2006) [29]	LSA, PLSA > LDA
	Anti-Phishing	
	• (Bergholz et al. 2008) [6]	Class-Topic Model
	Automatic Labeling	
	• (Magatti et al. 2009) [36]	Topics extracted by LDA
	Emotion topic	
	• (Bao, 2009) [4]	Social emotion
	Expert identification	
	• (Kongthon et al. 2009) [32]	Experts for R&D
	Role discovery	
	• (McCallum et al. 2007) [37]	Enron and academic email
	Sentiment summarization	
	• (Titov and McDonald 2008) [52]	Multi-Grain LDA
	Word sense disambiguation	
	• (Boyd-Graber et al. 2007) [12]	LDA with WordNet
CTM	Query classification	
	• (Zhai, 2009) [56]	Regularized CTM
	Topic detection	
	• (Chang & Boyd-Graber 2009) [13]	Predictive power ≠ semantic meaning
	Image retrieval	
	• (Greif et al. 2008) [23]	Precision: LDA,PLSA>CTM,
	Tracking objects	Predictive: CTM > LDA, PLSA
	• (Rodriguez et al, 2009) [41]	Scene tracking

Note: 1) SHA for message-digest algorithm 5 and KNN for K-nearest neighbor algorithm.

the probabilities of word-document occurrences, *P(d,w),* can be represented with $P(d)\Sigma_{z\epsilon Z}P(w|z)P(z|d)$ or alternatively with $P(d,w) = \Sigma_{z\epsilon Z}P(z)P(w|z)P(d|z)$ with Bayes' rule. Using EM algorithm which is a general solution in estimating unknown parameters, PLSA estimates topic probabilities *P(z)*, document probabilities given topics *P(d|z)*, and word probabilities given topics *P(w|z)*.

In PLSA, the values in *P(d|z)*, *P(z)*, and *P(w|z)* are interpreted as probabilities, but in LSA, the values just show loading values which have sometimes negative values. PLSA generally capture general themes or trends in documents. Hofmann [25] and Hofmann et al. [27] apply PLSA to identify factors to the *neural* journal and information retrieval for the first time. Other applications and performance of PLSA are shown in the third row of Table 1.

For synonyms, words in a topic from PLSA are more closely related than words in a topic from LSA. For polesemy, words in a topic from PLSA can be appeared in other topics simultaneously. One limitation of PLSA is that it does not consider documents generation, *P(d),* in that model. To fully reflect the generative process at a document level, latent Dirichlet allocation has been developed and that model is the topic of the next section.

3.3. Latent Dirichlet allocation

While addressing the limitation of PLSA, latent Dirichlet allocation (LDA) incorporates the generative process of documents with Dirichlet distribution. According to LDA process, each document is generated the following three steps. First, the number of words used in a document is determined by sampling with the Poisson distribution. Second, a distribution over topics for a document is elicited from the Dirichlet distribution. Third, based on the document-specific distribution, topics are generated, and then words for each topic are generated.

Like PLSA, LDA also provides topics in which words have probability values. LDA is appropriate for modeling long-length documents that contain multiple topics. For synonymy, words in a topic have similar meanings, but, interestingly, words are composed of adjectives and nouns like "good screen." This adjective and noun structure enables researchers to label topics with easy. LDA has been applied in many areas such as topic detection, emotion detection, and word sense disambiguation, and the fourth row in Table 1 shows applications and performance of LDA. For polysemy, words in a topic can be appeared in other topics simultaneously. However, LDA does not consider topic relations among topics. By identifying topic relations, we are able to understand deeply structures of documents.

3.4. Correlated topic model

The correlated topic model (CTM) has been developed to address the limitation of LDA. CTM can explicitly address topic relations using logistic normal distribution. In LDA, it is assumed that word in topics comprises a multinomial distribution, and multiple topics can be appeared in a document and the proportion of topics varies according to Dirichlet distribution. CTM follows the same generative process of LDA, but the difference is in the use of logistic normal distribution rather than Dirichlet distribution to capture topic relations.

In the Dirichlet distribution, the components of the proportions are practically independent, thus topics cannot have a relation with other topics. This independence prevents occurrence of a word in other topics. In fact, if a topic does not have a relation with other topics, a word in the topic also cannot be appeared in other topics. In order to reflect topic relation, CTM employs more flexible logistic normal distribution by considering covariance structure among the components of proportions. CTM are also used for many areas such as topic detection and image retrieval, and the last row of Table 1 shows applications and performance of CTM.

For synonyms, words in a topic have similar meaning, and those words are also similar with words of similar topics. For polesemy, words with different meanings can be shown up in other topics simultaneously. CTM paves the way for disambiguating senses of words. For example, because *programming java* topic has closer relation to *web service* topic rather than *coffee* topic, a word can be disambiguated with other similar topics. The characteristics and limitations of four text mining methods are summarized in Table 2.

3.5 The theoretical understanding for performance differences of four methods

The four text mining methods are fundamentally based on high-order co-occurrences and transitive relation [33]. The big difference among four methods is in that which parts of the structures each method exploits and how they capture the structure. In large, LSA captures unique structure but others seize overlapped structure. For instance, Figure 1 shows three documents where D_1 mentions (A, B, C), D_2 mentions (B, D, E), and D_3 mentions (D, F, G). In terms of documents, there is first order co-occurrence between D_1 and D_2 with B, and between D_2 and D_3 with D. There is second order co-occurrence between D_1 and D_3 because of transitive relation (A-B-D-F). Theses high-order co-occurrences become semantic structure of four text mining methods. While LSA captures unique parts (A,C – E – F,G) based on high order (B,D) first, generative models captures shared parts (B - D) first. Kontostathis and Pottenger find that second order co-occurrence of words has strongly affects to the results of LSA [33]. Thus, when applying four methods, researchers frequently separate documents into paragraph level in order to enrich co-occurrence patterns. Therefore, the performances of all four text

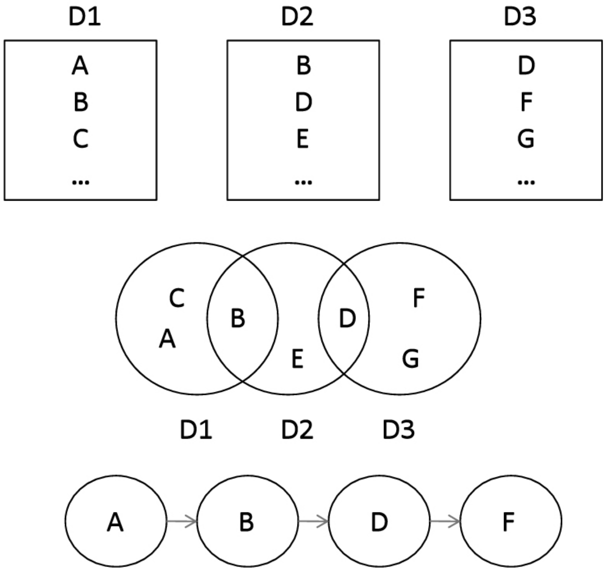

FIGURE 1. Co-occurrence structure of words among documents.

mining methods are fundamentally affected by high-order co-occurrences and transitive relation.

In LSA, the high order structure is captured with singular value decomposition with an implicit way. However, in generative models, the structure is captured explicitly with conditional independence [31]. The co-occurrence structure in Figure 1 can be expressed with conditional probabilities, *P(B|A), P(D|B),* and *P(F|D)*, showing A-B-D-F relation. The versatile of generative models arises from modeling co-occurrences structure explicitly. PLSA, LDA, and CTM incorporate this prior knowledge to each model. Thus, statistically generative models have more flexibility than discriminative models.

TABLE 2. The characteristics and limitations of four text mining methods

Models	**Characteristics / Limitations**
Latent Semantic Analysis (LSA) [18, 34]	Characteristics • Reduces dimensionality of *tf-idf* using Singular Value Decomposition. • Captures synonyms of words. • Not robust statistical background. Limitations • Difficult to determine the number of topics. • Difficult to interpret loading values with probability meaning. • Difficult to label a topic in some cases using words in the topic.
Probabilistic Latent Semantic Analysis (PLSA) [19-20, 25, 39]	Characteristics • Mixture components are multinomial random variables that can be viewed as representations of "topics." • Each word is generated from a single topic; different words in a document may be generated from different topics. • PLSA partially handles polysemy. Limitations • No probabilistic model at the level of documents
Latent Dirichelet Allocation (LDA) [10, 22]	Characteristics • Provides full generative model with multinomial distribution for words in topics and Dirichlet distribution over topics. • Handles long-length documents. • Shows adjectives and nouns in topics. Limitations • Incapable to model relations among topics.
Correlated Topic Model (CTM)[8]	Characteristics • Considers relations among topics using logistic normal distribution • Allows the occurrences of words in other topics. • Allows topic graphs. Limitations • Requires complex computations. • Contains too general words in topics.

TABLE 3. The theoretically appropriate conditions for high performance.

Methods	**Conditions**
LSA	Documents have to contain many high-order co-occurrences and transitive relations. Throughout relations, LSA exploits unique structure as factors [33]. Papadimitriou et al. show that LSA can get good performance when documents have 1) no style modifier, 2) a single topic in a document, and 3) peculiar terms in one topic [39].
PLSA	Beside high-order co-occurrence and transitive relations, generative models require conditional independence. This can be checked with correlation analysis or separation in causal map [31]. If a length of a document is long, many words in the document probably have relation within the document but with other documents, which reduces high-order co-occurrences. Thus, long lengthy documents are not appropriate for PLSA.
LDA	Theoretically, LDA can handle mixed lengthy documents. However, because of exchangeability, words in a topic can be collected from any part in a document, and the meaning of topic will be ambiguous. To enrich transitive relations and high-order co-occurrence, in some cases, LDA are applied to many small documents that are obtained by dividing documents into paragraph level.
CTM	Topic relations can be extracted from many co-occurrence relations. Thus, documents in online discussion forum or QnA might produce meaningful relations among topics.

Because four models are based on high-order structure and transitive relation, the examination of two assumptions is crucial to get high performance. These assumptions can be checked by investigating co-occurrence through reading texts, examining frequencies of words, and making a correlation matrix with respect to documents. Additionally, conditional independence for generative models can be checked by correlation analysis and separation in causal maps [31]. If text data does not have such co-occurrence structure or conditional independence, the performance of text mining can be deteriorated. However, in spite of various tests, no tests are definitive to predict performance. Table 3 presents some guidelines to apply four text mining methods.

4. EMPIRICAL ANALYSIS

We present two examples for the empirical comparison of four text mining methods; topic detection and spam filtering. In topic detection, we examine the characteristics and limitations of four methods with a descriptive way, but in spam filtering, we compare them with an objective performance measures.

4.1 An illustrative example of topic detection

Topic detection is a general application of almost all text mining approaches from LSA to CTM. When the proponents of each model propose their models in their studies, their usual application of the models is on topic detection. Because all four approaches can be applied to the topic detection study directly, an experiment using the same documents can reveal similarity and difference among four models. Moreover, to our knowledge, because there has been no such experiment of the comparison, this analysis might shed light on the clarification of fortes and limitations of each approach. However, because topics is not *definitive* but subjective [9], we compare the performance of topic detection with a descriptive way.

For the experiment, we specifically collect users' comments on a camera (Canon PowerShot A590IS 8MP Digital Camera with 4x Optical Image Stabilized Zoom) from Amazon website because users' comments on a product can be not only valuable guidance for developing new products but also valuable materials for adjusting defects of products to companies. We collect 258 comments for the camera but exclude 4 comments because the reviews are written with Spanish language. By applying preprocessing of text mining such as excluding of stopwords, we obtain a 254-3,189 dimensions of a document-by-word matrix. Then, by fixing the number of topics with 10, we apply four text mining methods to this matrix.

TABLE 4. The four documents of the topic #3

Docs	Contents
#24	I gave this camera to my son and his **fiance** and they simply **love** it. It's not too **complicated**, has lots of features and takes great photos.
#70	**Love** it so much that I bought a second one for my **nephew**. Steve of **StevesDigicams**.com loved it too.
#68	I just bought this Cannon and I **love** it. I **recomend** it for sure. Great quality photos and very **proffesional**.
#7	I can even take close up photos of **books** to show on computer. **Love** the movie mode, too.

Table 5 shows topic solutions that are obtained by applying four text mining methods. The topics of LSA have two characteristic; distinctive and peculiar words *within* a topic and exclusive topics *with* other topics. These characteristics can be attributed to both LSA itself and co-occurrence structure. In the model side, LSA is just dimensional reduction without considering any meaning, and inherently orthogonal between topics. In the co-occurrence structure side, LSA captures unique words. Table 4 shows second-order structure of documents for topic 3. In those documents, *love* is shared by four documents, but document #24 has *fiance* and *complicated* and document #68 has *recommend* and *proffessional*. Because of the shared word, *love,* four documents belong to the topic 3, but because of peculiar words, *fiance* and *recommend*, the topic 3 also has unique words in each document.

PLSA shows three distinct characteristics. First, product name and its accessories have high probabilities. Second, *camera* is shown up over several topics simultaneously, so that polysemy issue is touched although the meaning of it is not clear. At least, we can think that while *camera* in topic 1 is about photo or pictures, *camera* in topic 2 is about battery. Third, topics are about general theme or trend. For example, we can think that topic 1 is about *photo* theme and topic 2 is about *battery* theme. These characteristics can be contributed to PLSA itself and co-occurrence structure. In the model side, a word having high marginal probability will be used firstly, and product name or accessories can have high marginal probabilities because of frequent uses. In the co-occurrence structure, PLSA captures shared words instead of peculiar words in each document.

TABLE 5. Topic solutions for four text mining methods.

Method	Topic	Words
LSA	#2	play, pentax, operates, stop, company, wish, comparing, array, cards, mentioned[1)]
	#3	fiance, proffessional, recomend, nephew, stevesdigicams, disposable, fail, complicated, love, books[2)]
PLSA	#1	camera, great, pictures, digital, features, canon, cameras, easy, point, good, quality, photos, price
	#8	camera, batteries, good, photos, just, battery, shots, took, great, like, time used, light
LDA	#1	camera, batteries, good, battery, cannon, pictures, just, want, easy, like, life, shots
	#2	camera, flash, good, mode, zoom, canon, picture, pictures, better, manual, time, digital, great
CTM	#0	camera, image, great, powershot, shot, product, digital, good, like, zoom, optical, amazon
	#5	camera, canon, image, cameras, noise, viewfinder, digital, shoot, pictures, quality, great, point, uses, good

Note: 1) Words are ordered according to high loadings values.
2) Words are ordered according to high probabilities.

LDA shows three distinct characteristics. First, like PLSA, topics in LDA are composed of product name and accessories. This is happened with high marginal probabilities of those words. Second, there are co-occurrences of adjective and noun, which allows for easy labeling. For example, topic 1 can be named with “good battery” because of close occurrences of two words. Third, LDA provides good modeling for long lengthy documents. For example, one of long lengthy document discusses about battery efficiency, picture quality, and services of company. These topics of the document are captured in corresponding topics of LDA. In terms of co-occurrence structure, we can conjecture that because quality of picture, price, and battery are main concerns of users, documents have high co-occurrence patterns for those words. In fact, we are able to find over 150 co-occurrences of those words.

Finally, the topic solution of CTM reflects the relations among topics as well as topics. Like LDA, CTM also has adjectives and nouns for each topic with high probabilities, and that allows easy labeling of topics. Moreover, CTM can identify relations among topics. For example, we are able to make 5 topic groups from 10 topics such as (0, 5), (1, 6), (2, 7), (3, 8), and (4, 9). For example, topic 0 and Topic 5 share *camera*, *image*, *digital*, and *good* words. Thus, CTM can be used to not only elicit topics but also to identify relations among topics. Table 6 show a summary of major findings from four text mining methods.

Overall, as a discriminative model, LSA focuses on uniqueness in each topic. Generative models such as PLSA, LDA, and CTM emphasizes general themes in documents. PLSA can produce best result in single topic for one document, and LDA and CTM produce best result regardless of single or multiple topics for one document. Moreover, CTM can model the relations between topics, and the relations can be used for further grouping or identifying the relations among topics.

4.2 Spam filtering

Spam mail has become an important issue in information technology because it wastes the time of users for deleting spam mails [46]. For solving this problem, various spam filtering methods such as machine learning, Bayesian classification, and statistical learning have been developed. Spam filtering can be regarded not only as text categorization problem for classifying spam and legitimate mail but also as information retrieval problem for selecting the highest matched mail between two types of mails. Also, it is connected to clustering problem for grouping mails. Because many text mining methods are employed for spam filtering, and the performance measurement is relatively objective, spam filtering could be another for performance comparison. For the comparison of four methods, we specifically pay an attention to single type of an application rather than a mixed method combined with text mining methods.

We use a corpus known as Ling-Spam corpus, where 2,893 mails to the Linguist mailing list were manually classified [3]. Among four separate corpus, we use the lemm corpus. Out of 2,893, there are 481 spam mails, which is about 16.6%. We use 10-fold cross-validation method that nine parts are used for training and one part is used for testing with 10 times. Each fold has the same size with 289 or 290 mails. In addition, we set the number of topics for four text mining methods with 30 topics.

LSA has been applied to spam filter for a long time [5, 21]. The spam filtering in LSA has two phases. In the training phase, singular value decomposition is applied to the term-document matrix that is obtained from a training data set. The average dimension of a term-document for training set is 39,992-2313. The classification phase is consisted of the query indexing with test mails and the retrieval of the closest mail from the training set. If the closest mail from the training set is a spam mail, it is classified as a spam mail; otherwise it is classified as a legitimate mail. Table 7 presents the performance of four text mining methods in terms of various evaluation metrics such as precision, recall, accuracy, and weighted accuracy (WA), and total cost ratio (TC). Intuitively, the error of classifying a legitimate mail as a spam mail is more dangerous than classifying a spam mail as a legitimate mail (a false positive). For incorporating asymmetric risk, we set the cost ratio with 9 for WA and TC with reasonable cases.

In precision and recall of spam, recall is greater than precision because false positive (LS: Legitimate -> Spam) is greater than false negative (SL) by 200%. The overall accuracy is 0.969, but with cost ratio 9, it is increased to 0.974 for weighted accuracy. In LSA, TC is 4.123, and higher TC indicates better performance.

PLSA can be applied to spam filtering using the same approach in LSA using information retrieval and text categorization

TABLE 6. The major characteristics of topics from four methods

Method	Description
LSA	• Distinctive and peculiar topic • Exclusive topics (no relation with other topic)
PLSA	• High probabilities of product name and parts (camera and batteries) • Polysemy (multiple occurrences of camera over topics) • General topic reflecting theme and trend
LDA	• High probabilities of product name and parts • The close co-occurrences of adjective and noun word (easy labeling) • Long lengthy document
CTM	• Adjective + noun structure (easy labeling) • Identifying relations among topics

TABLE 7. The performance of four text mining methods in spam filtering

	LSA	PLSA	LDA	CTM
Spam Precision[1]	0.881	**0.895**	0.892	0.891
Spam Recall[2]	0.942	**0.955**	0.949	0.949
F_1[3]	0.911	**0.924**	0.920	0.920
LS/Total	0.021	**0.018**	**0.018**	0.019
LS/SL	2.178	**2.524**	2.250	2.292
Accuracy[4]	0.969	**0.974**	0.973	0.972
Weighted Accuracy[5]	0.974	**0.978**	0.977	0.976
Total cost ratio[6]	4.123	**4.793**	4.688	4.597

Note: 1) Spam Precision = SS / (SS + SL), where SS denotes that a spam mail is classified as a spam mail, and LS denotes that a legitimate mail is classified as a spam mail. 2) Spam Recall = SS / (SS + LS). 3) F_1 = (2 x Precision x Recall) / (Precision + Recall). 4) Accuracy = (LL + SS) / (LL + SS + LS + SL). 5) With the cost ratio (CR) 9, weight accuracy = (CR x LL + SS) / (CR (LL + LS) + (SS + SL)). 6) Total cost ratio = (LL + SL) / (CR x LS + SL).

method [45]. We use the original PLSA model [26-27] instead of a sophisticated PLSA such as dual-PLSA [54]. By applying PLSA to the training set, we estimate parameters of the model and construct a lower-dimensional representation in the factor space. Then, test mails which were not part of a training set is folded in by fixing the *p(w|z)* and calculating weights *p(z|q)* through each M-step. The third column of Table 9 shows the performance of PLSA. Compared to LSA, PLSA shows better performance. The precision is increased about 2% and the total cost ratio is increased about 16%. This increased performance can be attributed to the reduced false positive rate.

LDA cannot be applied to spam filtering directly using information retrieval because of few studies in that area is known. Instead, we apply a multi-corpus LDA method for spam filtering [7]. We classify a training set into two classes including spam and non-spam, and then run LDA separately for each class. We combine the resultant topic collection and make inference for test mails. The fourth column of Table 9 shows the performance of LDA. Compared to PLSA, the precision, recall, and F_1 are decreased, but those performances are higher than LSA. The rate of false positive to total is the same with LDA, but false negative is increased. Overall, the performance of LDA is almost the same with PLSA, and there is only 2% decreasing of TC.

CTM is applied to spam filtering with the same way of LDA. The recall, precision, and F_1 are very close to those of LDA. Compared to LDA, the number of SS is increased but the number of LL is decreased, where SS denotes that a spam mail is classified as a spam mail. However, false positive and false negative are almost the same with LDA. Because of the reduced number of SS, TC is decreased about 2%. In terms of TC, among four text mining methods, PLSA shows the highest performance, and next to LDA, CTM, and LSA in order. Also, between discriminative models and probabilistic models, probabilistic models show better performance in terms of precision, false positive, and TC. The experiment suggests that generative models may be promising for spam filtering.

5. CONCLUSION

In this paper, we have discussed the characteristics and limitations the four text mining methods including LSA, PLSA, LDA, and CTM. Also, we have discussed the theoretical backgrounds of four methods. The discussed methods are tested in two text mining areas; topic detection and spam filtering. In the topic detection, online reviews or comments of a camera from Amazon website with a descriptive way, and tested in the area spam filter. In the topic detection, we find that LSA can be used to seize unique or distinctive topics, and best applied to documents in which each document deals with a single topic. Other three methods as generative models put on weights on general themes or overall trends. PLSA can be best applied to single topic documents, but unlike topics in LSA, topics have general or overall characteristics. LDA can be best applied to documents in which each document deals with multiple topics. CTM can be used to identify the relationships among topics as well as topic detections. In the spam filtering, we measure classification performance of each model. We find that PLSA shows the highest performance in terms of total cost ration, and then LDA, CTM, and LSA in the order. In particular, generative models show higher performance than a discriminative model.

Among our contributions, the following two points are worthy of attention. First, we review four methods that are generally used in text mining research in order to introduce the advanced methods and the appropriate use in adequate context. By reviewing four text mining methods in the order of development, we show the characteristics and limitations of each method. Second, we present two examples that compare features and performances of each method. The comparison highlights the characteristics of each method with detailed explanations, and shows performance difference among models. Thus, our comparison can provide guidance for researchers and practitioners who applying text mining methods.

Though these contributions are noticeable, we admit the following three limitations of our study. Besides future research directions are presented to supplement the limitations.

First, out study is limited to the review of statistical approach, and does not extend to the syntactical and morphological approach of natural language. Given the advancement in the natural language, it is meaningful to review this approach in another study. Second, when making an experiment using comments data of a camera product, we arbitrarily set the number of topics with 10 topics. If the optimal number of topics for each method will be different, the fixed 10 topics bring a biased comparison. Also, we set the number of topics with 30 for spam filtering. However, still the automatic determination of optimal topics is far beyond without human judgments, and is one of emerging research topics in text mining. Third, when evaluating the quality of topics, we do not compare the result with the manually classified topics, and only focus on the causes showing different results in each model. However, ultimately, topics have to be useful for information users, and this type of comparison can reveal the quality of each method. Fourth, in the spam filtering, we employ typical approach in each text mining method. Finally, we acknowledge that each method needs testing in more various tasks such as word sense disambiguation and objects identification in pictures.

6. REFERENCES

[1] Ahrendt, P., Goutte, C., and Larsen, J., “Co-occurrence models in music genre classification”, IEEE International workshop on Machine Learning for Signal Processing, 2005, 247-252.

[2] Aldous, D., “Exchangeability and related topics”, *Ecole d’Ete de Probabilites de Saint-Flour XII, Springer Lecture Notes in Mathematics*, 1117, 1985, 1-198.

[3] Androutsopoulos, I., Koutsias, J., Chandrinos, K., and Spyropoulos, C., “An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages”, *ACM New York, NY, USA*, 2000, 160-167.

[4] 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.

[5] Bellegarda, J., Naik, D., and Silverman, K., “Automatic junk e-mail filtering based on latent content”, 2003, 465-470.

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

[7] Bíró, I., Szabó, J., and Benczúr, A., “Latent dirichlet allocation in web spam filtering”, *ACM New York, NY, USA*, 2008, 29-32.

[8] Blei, D.M., and Lafferty, J.D., “A Correlated Topic Model

of Science", *The Annals of Applied Statistics*, 1 (1), 2007, 17-35.

[9] Blei, D.M., and Lafferty, J.D., "*Topic Models*", Ashok N. Srivastava, Meharn Sahami ed., CRC Press, 2009.

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

[11] Bosch, A., Zisserman, A., and Munoz, X., "Scene classification via pLSA", *Lecture Notes in Computer Science*, 3954, 2006, 517-530.

[12] Boyd-Graber, J., Blei, D., and Zhu, X., "A topic model for word sense disambiguation", 2007, 1024-1033.

[13] Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D., "Reading Tea Leaves: How Humans Interpret Topic Models", *Neural Information Processing Systems*, 2009, 1-9.

[14] Chen, Q., Tai, X., Jiang, B., Li, G., and Zhao, J., "Medical Image Retrieval Based on Latent Semantic Indexing", Proceedings of the 2008 International Conference on Computer Science and Software Engineering, *IEEE Computer Society*, 2008, 561-564.

[15] Cheung, K., Kwok, J.T., Law, M.H., and Tsui, K., "Mining customer product ratings for personalized marketing", *Decision Support Systems*, 35, 2003, 231-243.

[16] Chou, T.-C., and Chen, M.C., "Using Incremental PLSI for Threshold-Resilient Online Event Analysis", *Knowledge and Data Engineering, IEEE Transactions on*, 20 (3), 2008, 289-299.

[17] Das, S.R., and Chen, M.Y., "Yahoo! for Amazon: Sentiment extraction from small talk on the web", *Management Science*, 53 (9), 2007, 1375-1388.

[18] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R., "Indexing by Latent Semantic Analysis", *Journal of the American Society for Information Science*, 41 (6), 1990, 391-407.

[19] Ding, C.H.Q., "A probabilistic model for Latent Semantic Indexing: Research Articles", *Journal of the American Society for Information Science and Technology*, 56 (6), 2005, 597-608.

[20] Fuhr, N., "Probabilistic models in information retrieval", *The Computer Journal*, 35 (3), 1992, 243-255.

[21] Gansterer, W., Janecek, A., and Neumayer, R., "Spam filtering based on latent semantic indexing", *Survey of Text Mining II: Clustering, Classification, and Retrieval*, 2008, 165-183.

[22] Girolami, M., and Kaban, A., "On an equivalence between PLSI and LDA", *ACM New York, NY, USA*, 2003, 433-434.

[23] Greif, T., Hörster, E., and Lienhart, R., "Correlated topic models for image retrieval", *Technical Report TR2008-09, University of Augsburg*, 2008.

[24] Herdiyeni, Y., Nurdiati, S., and Daud, I.A., "Image Semantic Extraction Using Latent Semantic Indexing on Image Retrieval Automatic-Annotation", Proceedings of the 2009 International Conference of Soft Computing and Pattern Recognition, *IEEE Computer Society*, 2009, 283-288.

[25] Hofmann, T., "Probabilistic latent semantic indexing", SIGIR-99, *ACM New York, NY, USA*, 1999, 50-57.

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

[27] Hofmann, T., Puzicha, J., and Jordan, M., "Unsupervised learning from dyadic data", *Advances in Neural Information Processing Systems*, 11, 1999.

[28] Ide, N., and Veronis, J., "Introduction to the special issue on word sense disambiguation: the state of the art", *Comput. Linguist.*, 24 (1), 1998, 2-40.

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

[30] 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.

[31] Koller, D., and Friedman, N., "*Probabilistic Graphical Models: Principles and Techniques*", The MIT Press, 2009.

[32] Kongthon, A., Haruechaiyasak, C., and Thaiprayoon, S., "Expert Identification for Multidisciplinary R&D Project Collaboration", *PICMET 2009 Proceedings*, 2009.

[33] Kontostathis, A., and Pottenger, W., "A framework for understanding Latent Semantic Indexing (LSI) performance", *Information Processing and Management*, 42 (1), 2006, 56-73.

[34] Landauer, T.K., Foltz, P.W., and Laham, D., "An introduction to latent semantic analysis", *Discourse processes*, 25, 1998, 259-284.

[35] Larsen, K.R., Monarchi, D.E., Hovorka, D.S., and Bailey, C.N., "Analyzing unstructured text data: using latent categorization to identify intelectual communities in information systems", *Decision Support Systems*, 45, 2008, 884-896.

[36] Magatti, D., Calegari, S., Ciucci, D., and Stella, F., "Automatic Labeling Of Topics", *Ninth International Conference on Intelligent Systems Design and Applications*, 2009, 1227-1232.

[37] 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.

[38] Mølgaard, L., Larsen, J., and Goutte, C., "Temporal analysis of text data using latent variable models", IEEE International Workshop on Machine Learning for Signal Processubg, 2009.

[39] Papadimitriou, C.H., Raghavan, P., Tamaki, H., and Vempala, S., "Latent semantic indexing: A probabilistic analysis", *Journal of Computer and System Sciences*, 61, 2000, 217-235.

[40] Pons-Porrata, A., Berlanga-Llavori, R., and Ruiz-Shulcloper, J., "Topic discovery based on text mining techniques", *Information Processing and Management*, 43 (3), 2007, 752-768.

[41] Rodriguez, M., Ali, S., and Kanade, T., "Tracking in Unstructured Crowded Scenes", *The 12 IEEE International Conference on Computer Vision*, 2009.

[42] 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.

[43] Rosenfeld, R., "Two decades of statistical language modeling: where do we go from here?", *Proceedings of the IEEE*, 88 (8), 2000, 1270-1278.

[44] Salton, G., "*Automatic text processing: the transformation, analysis, and retrieval of information by computer*", Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1989.

[45] Santhiappan, S., Gopalan, V.P., and Valarmathi, B., "Topic models based personalized spam filter", Proceedings of ISCF, 2006, 199-203.

[46] Sanz, E.P., Hidalgo, J.M.G., and Perez, J.C.C., "Email Spam Filtering", *Advances in Computers*, 74, 2008, 45-109.

[47] Sidorova, A., Evangelopoulos, N., Valacich, J., and Ramakrishnan, T., "Uncovering the intellectual core of the information systems discipline", *MIS Quarterly*, 32 (3), 2008, 467-482.

[48] Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., and Freeman, W.T., "Discovering objects and their location in images", International Conference on Computer Vision (ICCV 2005), 2005.

[49] Strunk Jr, W., "*The elements of style*", Filiquarian Publishing, LLC., 2007.

[50] Sun, J., Zhang, Q., Yuan, Z., Huang, W., Yan, X., and Dong, J., "Research of Spam Filtering System Based on LSA and SHA", *Springer*, 2008, 340.

[51] Tetlock, P.C., Saar-Tsechansky, M., and Macskassy, S., "More than words: Quantifying language to measure firms' fundamentals", *Journal of Finance*, 63 (3), 2008, 1437-1467.

[52] Titov, I., and McDonald, R., "A joint model of text and aspect ratings for sentiment summarization", *Urbana*, 51, 2008, 308-316.

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

[54] Xu, W., Liu, D., Guo, J., Cai, Y., and Hu, R., "Supervised Dual-PLSA for Personalized SMS Filtering", *Springer*, 2009, 254-264.

[55] Yang, W., and Dia, J., "Discovering cohesive subgroups from social networks for targeted advertising", *Expert Systems with Applications*, 34, 2008, 2029-2038.

[56] Zhai, H., Guo, J., Wu, Q., Cheng, X., Sheng, H., and Zhang, J., "Query Classification Based on Regularized Correlated Topic Model", 2009.