Information Processing and Management 52 (2016) 430–445

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Semantic search for public opinions on urban affairs: A probabilistic topic modeling-based approach

Baojun Ma [a], Nan Zhang [b,*], Guannan Liu [c], Liangqiang Li [d], Hua Yuan [d]

[a] School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, PR China
[b] School of Public Policy and Management, Tsinghua University, Beijing 100084, PR China
[c] School of Economics and Management, Tsinghua University, Beijing 100084, PR China
[d] School of Management and Economics, University of Electronic Science and Technology of China, Chengdu 611731, PR China

ARTICLE INFO

Article history:
Received 26 December 2014
Revised 12 September 2015
Accepted 14 October 2015
Available online 14 November 2015

Keywords:
Probabilistic topic modeling
Public opinions
Big data analysis
Semantic search
Latent Dirichlet allocation (LDA)

ABSTRACT

The explosion of online user-generated content (UGC) and the development of big data analysis provide a new opportunity and challenge to understand and respond to public opinions in the G2C e-government context. To better understand semantic searching of public comments on an online platform for citizens' opinions about urban affairs issues, this paper proposed an approach based on the latent Dirichlet allocation (LDA), a probabilistic topic modeling method, and designed a practical system to provide users—municipal administrators of B-city—with satisfying searching results and the longitudinal changing curves of related topics. The system is developed to respond to actual demand from B-city's local government, and the user evaluation experiment results show that a system based on the LDA method could provide information that is more helpful to relevant staff members. Municipal administrators could better understand citizens' online comments based on the proposed semantic search approach and could improve their decision-making process by considering public opinions.

1. Introduction

The spread of Web 2.0 applications enables a new model for the use of the Internet (Burke, 2009). Users act not only as websites' visitors but also as content creators (Ingawale, Dutta, Roy, & Seetharaman, 2013). User generated content (UGC) significantly enriches the amount of online information and makes it more difficult for users to comprehensively understand information through regular reading behavior (Li et al., 2012; Zhu, Mo, Wang, & Lu, 2011). However, it is obvious that to appreciate the significance of UGC, any use of the knowledge underlying UGC must be developed based on a comprehension of its content (Williams, Wiele, Iwaarden, & Eldridge, 2010).

With respect to public administrators, the Web 2.0 environment with multiple exchanges provides a vital opportunity for enhancing interactions between the government and citizens (Horton, 2006). The explosion of online UGC and the development of big data analysis provides a new opportunity and challenge to understand and respond to public opinions (Rhoda & Norman, 2013; Yu-Che & Tsui-Chuan, 2014). In particular, when confronted with urban residents whose problems are comparatively homogeneous, local governments attempt to establish interactive platforms to collect citizens' opinions and recommendations from various perspectives to serve as a foundation for evaluating the performance of the government and to guide future policy

* Corresponding author. Tel.: +86 10 62772746; fax: +86 10 62772746.
E-mail addresses: mabaojun@bupt.edu.cn (B. Ma), nanzhang@mail.tsinghua.edu.cn (N. Zhang), guannliu@gmail.com (G. Liu), langmalee@gmail.com (L. Li), yuanhua@uestc.edu.cn (H. Yuan).

http://dx.doi.org/10.1016/j.ipm.2015.10.004
0306-4573/

adjustment (Hong, 2013). However, as citizens' enthusiasm for voicing their opinions on the Internet grows, how to understand the information both timely and effectively becomes more significant, which is directly related to whether issues that concern administrators can be addressed and whether the corresponding feedback could be timely offered to citizens (Linders, 2012). Semantic analysis and semantic search technology are likely to become effective tools for assisting the public administrators in the rapid and precise positioning of mass text information.

This study focuses on a platform for acquiring online citizen opinions on urban public affairs issues. B-city has been honored as one of China's international metropolises. To reinforce communication between municipal administrators and citizens, to listen to citizens' opinions and suggestions on urban public affairs in a timely fashion, and to facilitate public participation in urban construction and development, in 2005 B-city's municipal government installed an opinion acquisition module related to urban public affairs onto its official website; the city now receives around 30,000 text messages per year.

Confronted with this non-structured text information, it is difficult to use simple statistical methods and traditional data processing tools to help officials better understand those comments. Before the start of the study, face to average more than 100 comments of daily public feedbacks, the office of the website, only sorts responsibility to the different departments manually, and oversees the feedbacks every day, without any historical data analysis tools. When sometimes need to summarize a period of public opinions, or analyze deeply around one case or one area, the office can only conduct keywords retrieval then manual read and summarize the retrieved results. Obviously, the current manual method is inefficient, even invalid for those tasks which unable to provide accurate keywords initially. Semantic analysis and semantic search could play significant roles in solving this problem.

The paper describes a basic idea for designing a semantic search tool directed to special demands: a framework composed of two sets of procedures—namely, a user search process and a probabilistic topic modeling process—is proposed based on the systemization of literature related to semantic search and semantic analysis to characterize the actual requirements of the online citizen opinion platform. In other words, the foreground procedure facilitates obtaining keywords from search input, provides auxiliary keywords to help searchers determine the theme (when necessary), and derives not only search results but also longitudinal changing curves. The background procedure is a process of preprocessing subject clustering for the comment data based on a probabilistic topic modeling approach called Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). Each theme that is generated based on probability subject modeling is appropriately viewed as the basic message block that waits constantly to be searched and invoked by the front-end search flow in the database. This study tries to make two aspects of contributions: (1) For this specific practice scenario mentioned in the paper, we provide a set of feasible solution to help the office search the public comments history data according to changeful requirements. In particular, the solution of the study is based on semantics, rather than merely based on keywords, which make the system "smarter" and could adapt to more complex requirements; (2) For academic area, we show a semantic search approach based on the LDA method. Compare to existing research involves the field of semantic modeling (Misra, Yvon, Cappé, & Jose, 2011; Xu, Zhang, & Wang, 2015), we try to show that the LDA could also play a role in the semantic search, and provide some key details of the techniques process including coordination between real-time search and non-real-time LDA calculation, keywords matching and suggestions in the user search process, determination of the number of latent topics.

To validate the usefulness and effectiveness of our proposed semantic search method based on LDA, we conduct a user evaluation experiment to compare with a baseline method, the Keywords-Matching approach (KM). The paper also presents the implications of the results for municipal management.

2. Brief literature review of semantic search

The idea of semantic search, which is understood as searching by meanings rather than literal strings of word and which aims to solve the limitations of keyword-based search models, has been the focus of a wide body of research in both the Semantic Web (SW) and the Information Retrieval (IR) communities (Fernández et al., 2011). An important aspect of semantic search approaches is that almost all of them use conceptual representations of content beyond mere keywords, and many of them also attempt to provide conceptual representations of user needs, as a method of enhancing traditional mainstream keyword-based search technologies.

Early in 2005, Mäkelä (2005) described five of the most-used methodologies in semantic search: (1) Resource Description Framework (RDF) Path Traversal; (2) keyword to concept mapping; (3) graph patterns; (4) logics; and (5) fuzzy concepts, fuzzy relations, and fuzzy logics. Then, Mangold (2007) surveyed and compared 22 different semantic search approaches or projects based on seven dimensions or criteria: architecture, coupling, transparency, user context, query modification, ontology structure and ontology technology. Thereafter, Dong, Hussain, and Chang (2008) conducted a brief survey based on a list of semantic search technologies from six categories: semantic search engines, semantic search methods, hybrid semantic search engines, semantic XML search engines, semantic ontology search engines and semantic multimedia search engines.

As described above, the core of semantic search technologies is the type and use of semantic knowledge representation. Accordingly, most of the semantic search methods in the literature can be distinguished according to the following three categories:

(1) *Linguistic conceptualization approaches* are based on light conceptualizations (usually considering few types of relationships among concepts) and low information specificity levels. For instance, early in 1998, Word Net was used to enhance search performance by considering the semantic relationships among words or concepts (Mandala, Takenobu, & Hozumi, 1998). Urbain, Goharian, and Frieder (2008) have explored unsupervised learning techniques for extracting semantic

information about biomedical concepts and topics and introduced a passage retrieval model based on Markov random fields for using these semantics in context to improve genomics literature searches.

(2) *Ontology-based proposals* consider a much more detailed and densely populated conceptual space in the form of ontology-based knowledge bases. For example, Chiang, Chua, and Storey (2001) have proposed a smart Web query (SWQ) method for the semantic search of Web data by utilizing domain semantics, which are represented as context ontologies to specify and formulate appropriate Web queries to search. Kim (2005) has designed and implemented an ontology-based Web retrieval system for semantic searches of the Web resources of international organizations such as the World Bank and the Organization for Economic Co-operation and Development (OECD). Moreover, the semantic Web has been widely used as an effective tool for ontology-based semantic search. For instance, Ding, Kolari, Ding, and Avancha (2007) have reviewed the methods and tools of using ontologies on the semantic Web, whereas Vandic, van Dam, and Frasincar (2012) have presented a platform for multifaceted product search using semantic Web technology. Although ontology-based semantic search approaches have been widely investigated and applied in recent years (Jiang & Tan, 2009; Johnson Lim, Liu, & Lee, 2010; Kara et al., 2012; Lee, Min, Oh, & Chung, 2014; Ruiz-Martínez, Valencia-García, Martínez-Béjar, & Hoffmann, 2012; Zheng, Chen, & Jiang, 2012), there are still various types of criticisms of their limitations, which primarily lie in the problem of knowledge incompleteness and obsolescence, along with the lack of established evaluation methodology for semantic search models (Blanco et al., 2013).

(3) *Statistical approaches*, such as LSA (Latent Semantic Analysis) (Papadimitriou, Tamaki, Raghavan, & Vempala, 1998), pLSI (Probabilistic Latent Sematic Index) (Hofmann, 1999b) and LDA (Blei, 2012; Blei et al., 2003), use statistical models to identify groups of words that commonly appear together and therefore may jointly describe a particular reality.

In addition, topic modeling algorithms—primarily regarded as the core techniques in statistical semantic search approaches—are statistical methods that analyze the words of an original text to discover the themes or topics that run through it. They were originally proposed and applied in the field of information retrieval (Blei, 2012) and have been adopted and used in various domains such as recommender systems (Wang & Blei, 2011), social sciences (Ramage, Rosen, Chuang, Manning, & McFarland, 2009), social network analysis (Li et al., 2010), opinion mining (Rao, Li, Mao, & Wenyin, 2014), etc.

Without directly addressing the topics inherent in the documents, the TF-IDF schema (Robertson & Jones, 1976) and Vector Space Model (VSM) (Salton, Wong, & Yang, 1975) have provided a rough solution to describing and modeling documents and their content or topic similarities, with the disadvantage that documents with a similar context but different vocabulary will not be associated with each other (Liu, 2007). To address this problem, Latent Semantic Analysis (LSA) has been proposed to convert documents' high-dimensionality word-space representations into low-dimensionality vectors of the topics (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), in which topics can be obtained using singular-value decomposition (SVD). Using such a technique yielded some improvement over a TF-IDF baseline.

A closely related technique—pLSI (Hofmann, 1999a)—attempts to create a set of topics in a probabilistic framework (Papadimitriou et al., 1998). pLSA topics are probabilistic instead of the heuristic geometric distances in LSA. pLSA has been successfully used for information retrieval related to large collections because it does not need to run the expensive SVD operation.

LDA is a popular topic modeling tool, designed to learn a set of topics (word distributions) and to infer mixtures of those topics to create low-dimensionality representations of documents (Blei et al., 2003), which further refines the pLSI model within a Bayesian framework. LDA has been the most popular topic extraction algorithm in recent years, and it has been successfully used to characterize the content topics of a document or a collection (Blei, 2012; Carterette & Chandar, 2009; Liu & Turtle, 2013).

In recent years, LDA model has been applied in various aspects of domain of information retrieval and natural language processing, such as text segmentation (Misra et al., 2011), feature identification (Xu et al., 2015), text classification (Phan, Nguyen, & Horiguchi, 2008), etc. Meanwhile, there have been many research works based on various applications of the LDA model in other research fields, such as customer segmentation (Wu & Chou, 2011), telecommunications fraud detection (Olszewski, 2012; Xing & Girolami, 2007), software evolution and defect reporting (Linstead, Lopes, & Baldi, 2008; Somasundaram & Murphy, 2012; Thomas, Adams, Hassan, & Blostein, 2014) as well as bioinformatics domain of microarray expression profile analysis (Bicego et al., 2012), miRNA-mRNA interactions study (Liu et al., 2010), the classification of genomic sequences (La Rosa, Fiannaca, Rizzo, & Urso, 2014), etc. Moreover, the dynamic topic model has been proposed to discover the evolution of topics over time, whose assumption is that the topics are related to the time periods, and the meanings of the words may evolve over time (Blei & Lafferty, 2006). However, it remains difficult to find research papers applying the LDA method in the domain of public management and public affairs (Levy & Franklin, 2014). Furthermore, there is little work addressing stages of user interaction by utilizing and applying the results of LDA modeling. A detailed introduction to the LDA model will be presented in the following section.

3. A probabilistic topic modeling-based semantic search model

3.1. The LDA model

Before describing our proposed semantic search model based on probabilistic topic modeling, we first must introduce the fundamental principle of the LDA model. Latent probabilistic topic models are effective methods for extracting latent semantic information from text documents (Blei, 2012; Blei et al., 2003; Deerwester et al., 1990; Hofmann, 1999a). Among these models, LDA appears to be the most effective (Blei et al., 2003).

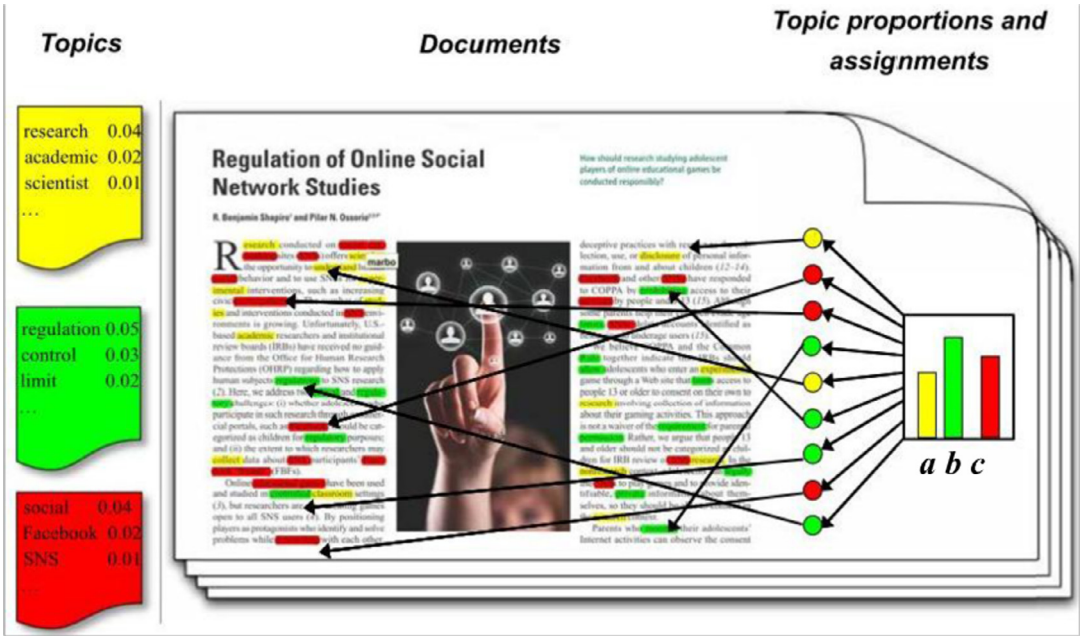

Fig. 1. The intuition underlying the LDA model.

The intuition underlying the LDA model is that documents have multiple topics (Blei, 2012). For instance, consider the paper in Fig. 1. This paper, recently published in Science Magazine and entitled "The Regulation of Online Social Network Studies", explores the US government's regulation of scientific research on online social networks, especially research related to adolescents (Shapiro & Ossorio, 2013). We highlighted, by hand, several words used in this article. Words about *scientific research*, such as "research" and "experiment", are highlighted in color ***a*** (i.e., yellow); words about *regulations*, such as "control" and "allow", are highlighted in color ***b*** (i.e., green); words about *social networks*, such as "SNS" and "Facebook", are highlighted in color ***c*** (i.e., red). Had we highlighted all of the words in the paper, we would have found that this paper blends scientific research, regulations and social networks in different proportions. Moreover, stop-words that make little topical sense—such as "and", "but" or "if"—have been excluded from this process. Furthermore, knowing that this paper blends those topics would help us to situate it in a collection of scientific or policy articles.

Before moving ahead, we first introduce the terminology and notations used in this model.

- A *word* $w \in \{1, \ldots, V\}$ is the most basic unit of discrete data. For clear notations, w is a V-dimensional unit-based vector. If w takes on the ith element in the vocabulary, then $w^i = 1$ and $w^j = 0$ for all $j \neq i$.
- A *document* is a sequence of N words denoted by $\mathbf{w} = (w_1, w_2, \ldots, w_N)$, where w_n is the nth word in the sequence.
- A *corpus* is a collection of M documents denoted by $D = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M)$.
- A *topic* $z \in \{1, \ldots, K\}$ is a probability distribution over the vocabulary of V words. Topics model particular groups of words that frequently occur together in documents, which thus can be interpreted as "subjects" or "themes". We denote $\mathbf{z} = (z_1, z_2, \ldots, z_N)$ as the sequence of topics across all words in a document.

LDA is an unsupervised, generative statistical model that seeks to capture the above intuition and proposes a stochastic procedure whereby the words in a document are generated. LDA is also a model of dimensional reduction, which handles the sparsity of vocabulary in a way that the words are represented as probability distribution over several hidden topics. The original space of vocabulary is mapped to several topics, and it can better depict the semantic meaning of the documents. Given a collection of unlabeled text documents, the LDA model seeks to discover hidden topics as distributions over the words in a fixed vocabulary. For example, the *regulations* topic has a high probability of containing words about regulations, whereas the *social networks* topic has a high probability of containing words about social networks. Here, words are modeled as *observed random variables* and topics are treated as *latent or hidden random variables*. Like other generative probabilistic modeling algorithms, once the generative procedure has been established, LDA first defines a joint distribution over both the observed and the hidden random variables and then utilizes statistical inferences to compute the conditional distribution or posterior distribution of the hidden variables (i.e., topics), given the observed variables (i.e., words).

In LDA, it is assumed that these topics are specified before any document has been generated. Thus, for any document in the corpus, the generative process contains two stages. First, a topic distribution vector modeled by a Dirichlet random variable (Balakrishnan & Nevzorov, 2005) has been chosen randomly to determine which topics are the most likely to appear in a document. Then, for each word that is to appear in the document, a single topic from the topic distribution vector is randomly selected. To actually generate the word, we then draw on the probability distribution conditioned on the chosen topic. Each word in

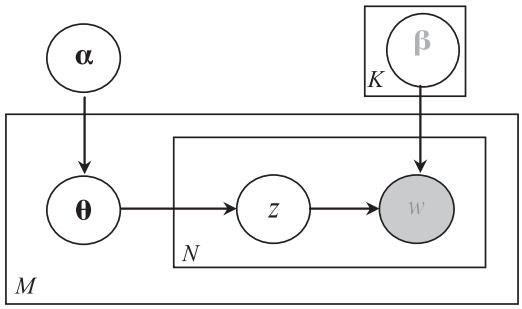

Fig. 2. Graphical model representation of LDA. The boxes are "plates" that represent replicates. The outer plate represents documents, whereas the inner plate represents the repeated choice of topics and words within a document.

a document is generated by a different, randomly selected topic. The above generative process is described more formally below (as shown in Fig. 2).

For each document indexed by $m \in \{1, \ldots, M\}$ in a corpus:

(1) Choose a K-dimensional topic distribution vector θ_m from the distribution $p(\theta|\alpha) = \text{Dirichlet}(\alpha)$.
(2) For each word indexed by $n \in \{1, \ldots, N\}$ in a document:
  (a) Choose a topic $z_n \in \{1, \ldots, K\}$ from the multinomial distribution $p(z_n = k|\theta_m) = \theta_m^k$.
  (b) Given the chosen topic z_n, select a word w_n from the conditional probability distribution $p(w_n = i|z_n = k, \beta) = \beta_{ik}$.

In the above descriptions, the parameter $\boldsymbol{\alpha}$ is a K-dimensional positive vector determining the Dirichlet distribution, which remains constant across all of the documents within a corpus. The Dirichlet distribution is given as follows:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta^{\alpha_i - 1} \tag{1}$$

where $\Gamma(x)$ is the gamma function.

The parameter $\boldsymbol{\beta}$ is a $V \times K$ matrix describing the word probabilities, which is also estimated from the data.

The generative procedure given above defines a joint distribution for each document. Assuming that parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are given, the joint distribution of a topic mixture is $\boldsymbol{\theta}$ and both a set of N topics $\mathbf{z}$ and a set of N words $\mathbf{w}$ is given by:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{2}$$

By integrating over $\boldsymbol{\theta}$ and summing over $\mathbf{z}$, we can obtain the marginal distribution or likelihood of a document:

$$\begin{aligned} p(\mathbf{w}|\alpha, \beta) &= \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta \\ &= \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{K} \prod_{k=1}^{V} (\theta_i \beta_{ik})^{w_n^k} \right) d\theta \end{aligned} \tag{3}$$

Normally, the optimal values of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are chosen to maximize the likelihood of all of the documents in the corpus:

$$l(\alpha, \beta) = \sum_{m=1}^{M} \log p(\mathbf{w}_m|\alpha, \beta) \tag{4}$$

In practice, the expectation–maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977; Hastie, Tibshirani, & Friedman, 2009) and the Gibbs-sampling-based algorithms (Casella & George, 1992; Porteous et al., 2008) are usually used to conduct the parameter estimation task.

More specifically, to estimate document-specific topic distribution probability (i.e., $p(z_n|\mathbf{w}_m)$) and topic-specific word distribution probability (i.e., $p(w_n|z_n)$), sample posterior distribution using Gibbs sampling is usually conducted. Given $\mathbf{z} = \{z_n = k, \mathbf{z}_{\neg n}\}$ and $\mathbf{w}_m = \{w_n = s, \mathbf{w}_{\neg n}\}$, it is easy to derive Eq. (5) as follows:

$$p(z_n = k|\mathbf{w}_m, \mathbf{z}_{\neg n}, \alpha, \beta) = \frac{p(\mathbf{z}, \mathbf{w}_m|\alpha, \beta)}{p(\mathbf{z}_{\neg n}, \mathbf{w}_m|\alpha, \beta)} \propto \frac{C_s^k + \beta_k}{\sum_{s=1}^{N} \left(C_s^k + \beta_k\right)} \times \frac{C_k^m + \alpha_k}{\sum_{k=1}^{K} \left(C_k^m + \alpha_k\right)} \tag{5}$$

where $\mathbf{z}_{\neg n}$ is the topic vector only excluded from z_n, $\mathbf{w}_{\neg n}$ is the word vector only excluded from w_n, C_s^k is the number of times word s is assigned to topic k and C_k^m is the number of times topic k is assigned to document m during the Gibbs sampling process, respectively.

Next, iterate Eq. (5) for all topics until it reaches the convergence for the entire document set. The final estimated results of $p(z_n|\mathbf{w}_m)$ and $p(w_n|z_n)$ are listed as follows:

- The document-specific topic distribution probability (i.e., $p(z_n|\mathbf{w}_m)$) can be derived as:

$$p(z_n|\mathbf{w}_m) = \theta_m^k = \frac{C_k^m + \alpha_k}{\sum_{k=1}^{K}\left(C_k^m + \alpha_k\right)} \tag{6}$$

- The topic-specific word distribution probability (i.e., $p(w_n|z_n)$) can be derived as:

$$p(w_n|z_n) = \frac{C_s^k + \beta_k}{\sum_{s=1}^{N}\left(C_s^k + \beta_k\right)} \tag{7}$$

From the above analysis, it is noted that the LDA model does not require any prior annotations or labeling of the documents and that the topics emerge from the analysis of the original texts. The LDA model enables us to organize and summarize electronic archives on a scale that would be impossible using human annotation (Blei, 2012).

3.2. Two-process semantic search framework

Unlike the traditional keyword-based or ontology-based IR models, we propose to design a novel semantic search framework for public comments on urban affairs based on a probabilistic topic modeling approach. More specifically, we propose to use LDA (Blei et al., 2003), which was introduced in detail in the last subsection. In this framework, the LDA model is utilized to discover and annotate large archives of public-generated documents with semantic thematic information to determine how these themes are connected to each other and how they change over time. The relation or association between a semantic topic and a document is what we call annotation. As introduced in the above subsection, document-specific topic distribution probability (i.e., $p(z_n|\mathbf{w}_m)$) can be derived by performing a Gibbs sampling.

As noted above, users—typically administrative staff of the website or a related government department—may be more interested in the longitudinal changing tendencies of particular topics of discussions, rather than merely a ranked list of documents that may be semantically related to a particular query. Therefore, in our proposed framework, after a user poses a query (a list of keywords or a sentence), our approach will ultimately generate and provide the end user with two types of results (i.e., the picture of the longitudinal changing curves of semantically related topics and a ranked list of documents for each semantically related topic).

In general, the overall semantic search model consists of the following two essential components (see Fig. 3), which are described in more detailed in the following subsections:

(1) The probabilistic topic modeling process (i.e., the LDA modeling stage); and
(2) The user search process.

3.3. Probabilistic topic modeling process

In this section, we will introduce detailed procedures for the probabilistic topic modeling stage by applying the LDA model.

In general, the probabilistic topic modeling process is based on (Chinese) text information in website documents that are generated by the public to provide comments or suggestions on diverse aspects of urban affairs and stored in the back-end database. To further explore that content, it is necessary to perform several pre-processing operations on the text, such as word segmentation and stop-word removal.

Because Chinese text does not encounter situations related to various syntactical forms (i.e., plural forms for nouns, gerund forms and past tense for verbs) (Liu, 2007) and the LDA modeling process below does not require the learning of each word's part-of-speech (Blei et al., 2003), it is noted that stemming and POS tagging are not necessary. To perform word segmentation on Chinese text, we choose the ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) system. The ICTCLAS Chinese word segmentation system is based on the Hierarchical Hidden Markov Model (HHMM) (Fine, Singer, & Tishby, 1998) and first proposed by Zhang, Yu, Xiong, and Liu (2003), which is one of the best and most widely used Chinese word-segmentation systems, reaching a segmentation precision of more than 97% (Du, Tan, Cheng, & Yun, 2010; Gao et al., 2013; Li, Ma, Zhang, Huang, & Kinshuk, 2013; Li, Zhu, & Zhou, 2014; Liu, He, Wang, Song, & Du, 2013; Rao et al., 2014; Shi & Nie, 2009). To complete the task of stop-word removal, we utilize a list of Chinese stop-words used in Apache Lucene's SmartChineseAnalyzer Class. Thereafter, each public comment document is represented as a vector of words.

Next, the core task lies in using the LDA model to conduct probabilistic topic modeling. Specifically, LDA modeling applies training and inference to all of the text vectors and can discover any latent topics or themes inherent in those data (Blei, 2012). In this paper, we actually choose the default parameter settings for the Dirichlet priors, i.e., we use symmetric setting $\alpha = 50/K$ and $\beta = 0.01$, where K is the number of topics. Such settings are widely used in previous literature (Wei & Croft, 2006; Zhang & Sun, 2012). Moreover, the Dirichlet priors would not influence the performance of the model much (Griffiths & Steyvers, 2004; Wei & Croft, 2006). We exploit Gibbs Sampling method to infer the parameters, which is in the form of Monte Carlo Markov Chain (MCMC) (Porteous et al., 2008). MCMC can be proved to get converged in multiple iterations of sampling process. To

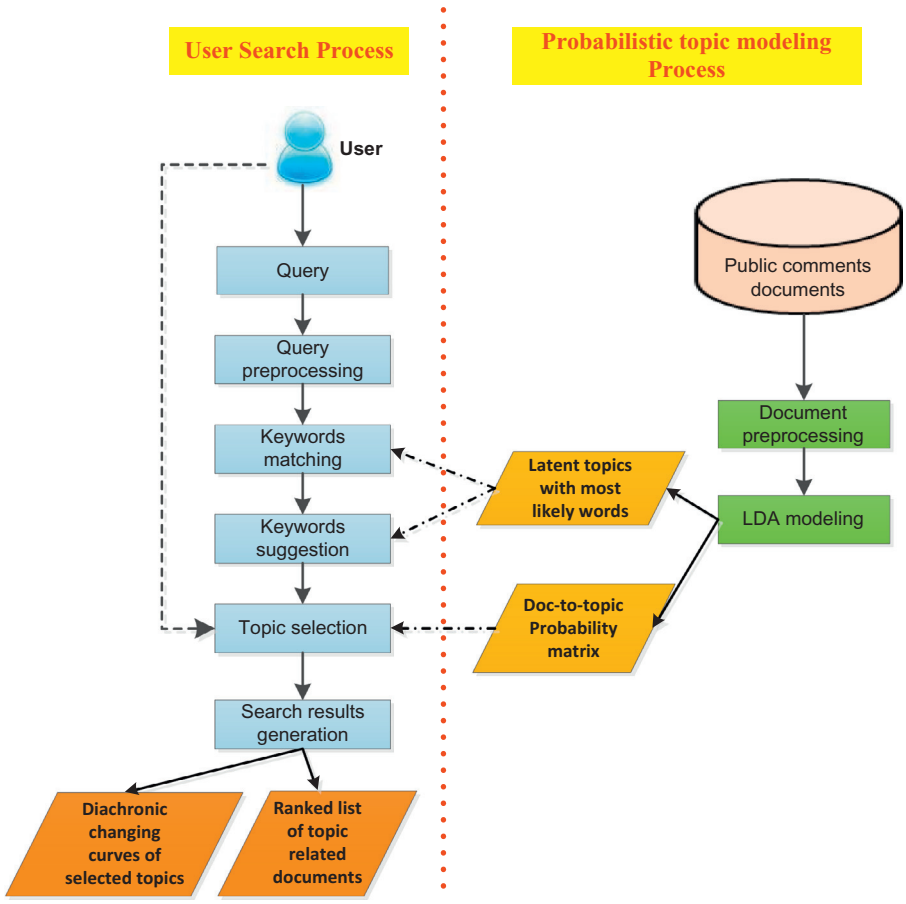

Fig. 3. Graphical representation of the proposed semantic search framework.

assess the convergence of the model, we computed the log likelihood in each iteration and the convergence was achieved when (old_likelihood-likelihood)/old_likelihood $<$ 10e$-$6.

After the topic modeling, by conducting the Gibbs sampling process it is easy to obtain the following useful results related to the words, documents and topics in the document set:

- Latent topics with the most likely words in each topic;
- A document-to-topic probability matrix.

Theoretically, for any given topic, there is a corresponding distribution across all of the words in the vocabulary (i.e., $p(w_t|z_n)$), and the LDA modeling process could provide the most likely words with the highest probability with respect to each topic, whereby the number of the most likely words can be given in advance.

In the document-to-topic probability matrix, each row corresponds to a document vector and each column represents a topic with $p(z_n|\mathbf{w}_m)$ in each cell, in which z_n is the topic k, $\mathbf{w}_m$ represents text document m and $p(z_n|\mathbf{w}_m)$ denotes the probability $\mathbf{w}_m$ belonging or discussing z_n.

A crucial issue in LDA modeling is to determine the number of latent topics K, which may impose an impact on the modeling results. Although several researchers have already suggested using the value of $p(\mathbf{w}|K)$ (Griffiths & Steyvers, 2004) or perplexity (Blei et al., 2003) to evaluate the topic modeling, to find the optimal topic number, it is still argued that the evaluation process often suffers from an over-fitting problem (Blei, 2012; Blei et al., 2003). In practice, we recommend using the 10-fold cross-validation of perplexity to avoid the possible over-fitting problem. The perplexity used by convention in language modeling, is monotonically decreasing in the likelihood of the test data, and is algebraically equivalent to the inverse of the geometric mean per-word likelihood (Blei et al., 2003), which is illustrated as:

$$perplexity(D_{test}) = \exp\left\{-\frac{\sum_{m=1}^{M}\log p(\mathbf{w}_m)}{\sum_{m=1}^{M} N_m}\right\} \tag{8}$$

where $\mathbf{w}_m$ denotes text document m and N_m represents the number of words in document m. Moreover, a lower perplexity score indicates better generalization performance.

Specifically, to conduct the 10-fold cross-validation process, the entire document dataset has been randomly partitioned into 10 equal sized subsamples. Of the 10 subsamples, a single subsample is retained as the test dataset for testing the model, and the remaining 9 subsamples are used as training dataset. The cross-validation process is then repeated 10 times (the folds), with each of the 10 subsamples used exactly once as the test dataset. The 10 results of perplexity from the folds can then be averaged to produce a single evaluation of modeling effectiveness. The optimal topic number can be determined by the lowest average value of perplexity.

Furthermore, given the topic number K, the total number of documents M, the average number of words in each document N and the maximum number of iterations of sampling process T, the time complexity for LDA Modeling by adopting Gibbs sampling is $O(TKMN)$ (Porteous et al., 2008). When referring to the vocabulary size V, though we need to sample the latent assignment for each word appearance (not the unique word) in each document in Gibbs Sampling implying $V < M*N$, it is easy to derive that the time complexity for LDA Modeling can also be represented as $O(TKV)$.

Obviously, with large values of T, M and V, the LDA modeling would be a relatively time-consuming process, which however would not pose great influence on user search process described in Section 3.4. Though the user search process must be conducted immediately after a user query is submitted, the LDA modeling process (including the determination of optimal topic number) can be conducted offline. By assuming that the topics of the public comments on urban affairs are relatively stable, we could perform the LDA modeling process each time after a fixed time period (e.g., each week) or after certain burst events appearing (e.g., new policy announced) when searching is idle, merely matching results when users conduct a real search, largely reducing the computation burden. During the period before new modeling conducted, the document-to-topic probability vectors of new generated public comments with relative small volume can be easily estimated by the offline inference process (Blei et al., 2003).

3.4. User search process

Based on the results of the periodic performance of LDA modeling, once a user has made a query request, the user search process described here could be designed to provide the end user with the results in which he or she is interested. The sub-steps involved in the user search stage are as follows:

Sub-step 1: Query preprocessing.
Sub-step 2: Keyword matching.
Sub-step 3: Keyword suggestion.
Sub-step 4: Topic selection.
Sub-step 5: Search results generation.

Normally, a user query appears as a list of several keywords or a complete sentence. In the latter situation, our approach would conduct text pre-processing operations similar to those in the above section, such as word segmentation and stop-word removal. Moreover, similar to the ontology-based method, normally only the words that possess actual meaning, such as nouns and verbs, will pose an important influence on semantics or meaning. Therefore, the part-of-speech (POS) tagging technique is used to label these keywords; meaningless words are then filtered (Allan & Raghavan, 2002; Toutanova & Manning, 2000). Next, after the user query is processed, a list of meaningful keywords from the query are used to conduct keyword matching based on the latent topics with the most likely words in each topic, which are generated by the LDA modeling process. Specifically, in this sub-step, we need to find all of the latent topics with thematic keywords that contain at least one of the meaningful keywords from the query.

After keyword matching, if none of the topics are selected, the user will be returned a notification of empty results and a suggestion for query modification. Otherwise, for each possible semantically related topic, the first several thematic keywords with the highest probabilities will be recommended to the user, respectively. Then, the user can read the recommended keywords for each possible semantically related topic and select the appropriate topic or topics of interest.

With the user-selected topics, based on the document-to-topic probability matrix and user-specific time period, two types of semantic results will be generated and provided to the end user: the picture of changing trends of semantic-related topics and a ranked list of documents for each semantic-related topic.

In more detail, after we obtain all of the user's topics of interest, certain related statistical information (such as related-topics discussion hotness per day) can be calculated and provided for the end user. Specifically, to show the discussion hotness of these semantically related topics each day, the overall relevance of the documents to the user-selected topics—i.e., the daily thematic cumulative relevance—can be calculated. In general, given the total document set D_i in day d_i and the topics selected by end user $\mathbf{z}_{CT}$, we can find the daily thematic cumulative relevance $CumRel(\mathbf{z}_{CT}, d_i)$ as follows. For each document $\mathbf{w}_i$, its relevance to $\mathbf{z}_{CT}$ can be regarded as the summation of the probability of $\mathbf{w}_i$ with respect to each of the selected topics, which can be easily derived from the topic-document distribution matrix:

$$Rel(\mathbf{z}_{CT}|\mathbf{w}_i) = \sum_{z_n \in \mathbf{z}_{CT}} p(z_n|\mathbf{w}_i) \tag{9}$$

Next, the daily thematic cumulative relevance $CumRel(\mathbf{z}_{CT}, d_i)$ is represented as the summation on the total document set D_i in day d_i:

$$CumRel(\mathbf{z}_{CT}, d_i) = \sum_{\mathbf{w}_i \in D_i} Rel(\mathbf{z}_{CT}|\mathbf{w}_i) = \sum_{\mathbf{w}_i \in D_i} \sum_{z_n \in \mathbf{z}_{CT}} p(z_n|\mathbf{w}_i) \tag{10}$$

Furthermore, the end user can limit the period of the search results to a particular interval, such as ***TI***, when he or she submits the semantic query request. Accordingly, when the above results are limited to period ***TI***, they can generate a picture of changing trends in semantically related topics.

In addition, by ranking all of the documents in period ***TI*** based on their thematic relevance to $\mathbf{z}_{CT}$, i.e., $Rel(\mathbf{z}_{CT}|\mathbf{w}_i)$ and by selecting the top k documents, we can also provide the end user with a ranked list of semantically related documents. In particular, we add the user interaction sub-step (Step 4) to enable users to express their search intentions more accurately and to achieve a more flexible semantic search.

4. Practical system implementation

This study's practical setting is an online platform to collect citizens' opinions on urban public affairs in B-city, one of China's international metropolises. To reinforce communication between municipal administrators and citizens, to pay timely attention to citizens' opinions and suggestions on urban public affairs, and to facilitate public participation in urban construction and development, in 2005 B-city's municipal government installed an opinion acquisition module related to urban public affairs on its official website. Since 2010, the module has been accessible through mobile terminals, in light of the popularity of mobile communications technology. Citizens can voice their opinions and recommendations about urban public affairs in the form of online comments at any time. Moreover, they can present comments and inquiries related to urban public service procedures along with reports and complaints about improper statements and actions by either government offices or by specific civil servants. The official website of B-city's municipal government notes the existence of a dedicated team responsible for processing citizen comments in a timely fashion, for allocating the items to related government departments according to their contents, and for supervising handling processes. Since it first went online, the platform has witnessed eight years of development, and its 240,000 documents of public comments constitute a set of large-scale data. A document here is defined as the text content a citizen submitted to the official website of B-city's municipal government as his or her opinions and recommendations about urban public affairs once.

Confronted with this non-structured text information, the platform's administrators believe that a large amount of valuable knowledge is contained therein—for example, the issues that are matters of concern for the general public and the variance in the hotness of those problems due to policy promulgation, department reorganization, or even environmental change. However, non-structured text information and ambiguous expressions render administrators unable to derive the rules underlying citizen comments. Although the department has established a five-person team to classify and generalize the comments, understanding them remains difficult. Semantic analysis and semantic search will play significant roles in solving this problem.

Therefore, B-city's municipal government, platform administrators and end users of the related information require a specific search tool that could meet the following requirements: (1) they need a system to conduct topic searches by keyword and to rapidly search a specific topic from the mass of comments; (2) in view of the difference in language expressions and wordings among various citizens, such a search should be able to implement the association of the meanings of citizen comments, rather than merely comparing whether the keywords appear; (3) with a view to the various meanings of some words, the system should be able to help searchers quickly choose the topics of concern by providing other auxiliary keywords when multi-vocal keywords appear; (4) comments that are subordinate to the hot topics should be retrievable, but it is also important to understand the longitudinal changing curves based on statistical data to measure and evaluate the occurrence and governance of the city's public affairs issues, or even the effectiveness of associated government sectors; and (5) because searches are directed to specific text databases with limited scales, the abovementioned objective must be achieved using fewer computing resource expenditures to their greatest ability and at an appropriate search speed.

In our practical system implementation, the whole process for a user's search can be illustrated with a sequence diagram as following in Fig. 4.

In such a search system, there are two main actors: the user and the administrator, whose use-case diagrams are shown in Fig. 5. The activity of an information query is launched by a user and three actions of keywords inputting, thematic keywords selecting and query resetting are involved (see Fig. 5(a)). As for the administrator, a special management activity is topic modeling, in which, two important resources of topic-keywords table and doc-to-topic matrix need to be maintained regularly (see Fig. 5(b)).

In our semantic search system, to find the "optimal" topic number for the 24,000 documents of public comments, we followed the 10-fold cross-validation process with topic number from 20 to 400 (each 20 as an interval) to obtain the curves of perplexity changes with number of topics, which is illustrated in Fig. 6. In Fig. 6, "Test 1" to "Test 10" represents the result when the corresponding partitioned subsample was treated as the test dataset and "Average" is the averaged result for the whole dataset. It is shown that in most cases, the values of perplexity for Test 1 to Test 10 reach relatively low scores when the topic number is in the range of 100–200. In addition, the Average curve gets its lowest point in the case that the number of topics is 140. Therefore, the "theoretical optimal" number of topics for LDA modeling of these 24,000 documents should be 140. In addition, the vocabulary size of this model, i.e., the number of unique words after data preprocessing, is 314,305.

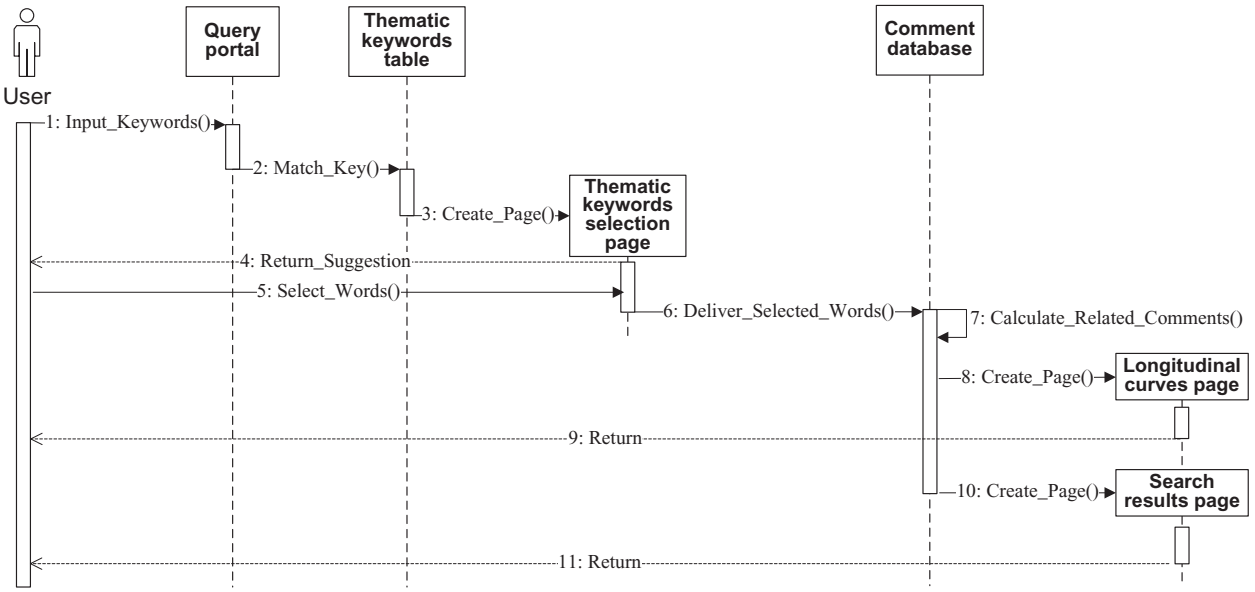

Fig. 4. The sequence diagram for the semantic search system.

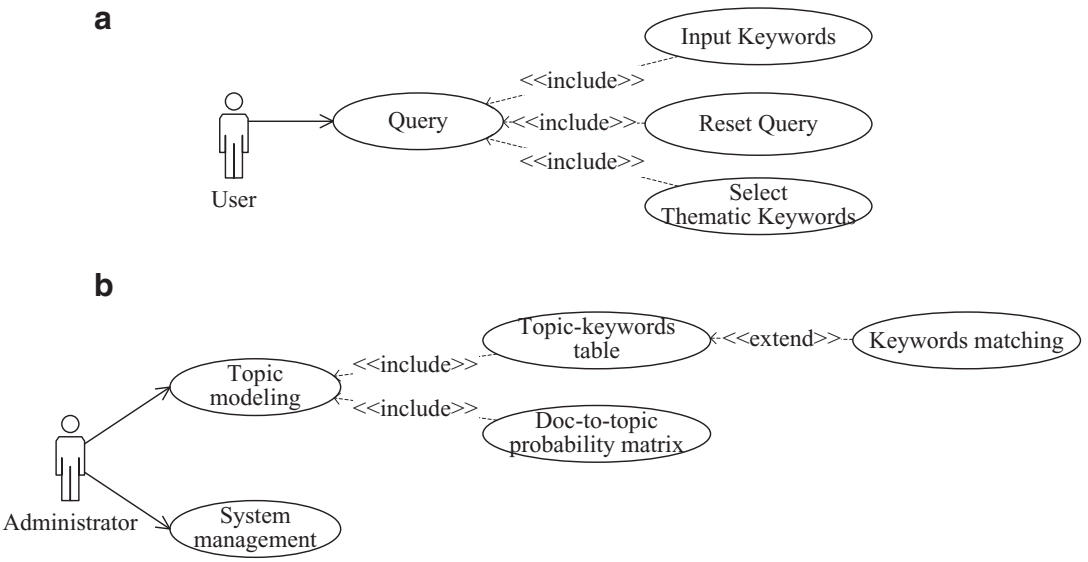

Fig. 5. The use-case diagrams for the user (*a*) and administrator (*b*) in the semantic search system.

Fig. 7 shows the interface of our search system based on the probabilistic topic modeling approach. When an end user submits a semantic query with the keyword "公交 (i.e., Public Transportation)" to the system, by matching the keyword "公交" with the thematic keywords of the latent topics extracted by the LDA model, seven latent topics, each with ten suggested keywords, have been recommended to the end user to ask him to select real interest subtopics semantically related to "公交".

Next, further assume that the end user is indeed interested in public comments and discussions about the topic of "bus IC card recharge" problem and that he selects Topic 5. Thus, our proposed approach would first provide a picture of the longitudinal changing curves of topic-discussion hotness for the corresponding topic (i.e., bus IC card recharge), which is shown in Fig. 8.

Moreover, based on the semantic relevance of each public comment with respect to the user-selected topic, i.e., bus IC card recharge, our proposed method could also provide the end user with a ranked list of public comments semantically related to the user-selected topic. Due to limited space, Fig. 9 only lists the title of each semantically related comment with its issue date and relevance score to corresponding topic.

The above words and figures show a complete semantic search process using an example. For users, the tool could provide an easy search that helps them to focus on certain topics of interest and to understand the latest hot topics; it also summarizes longitudinal changing trends over time.

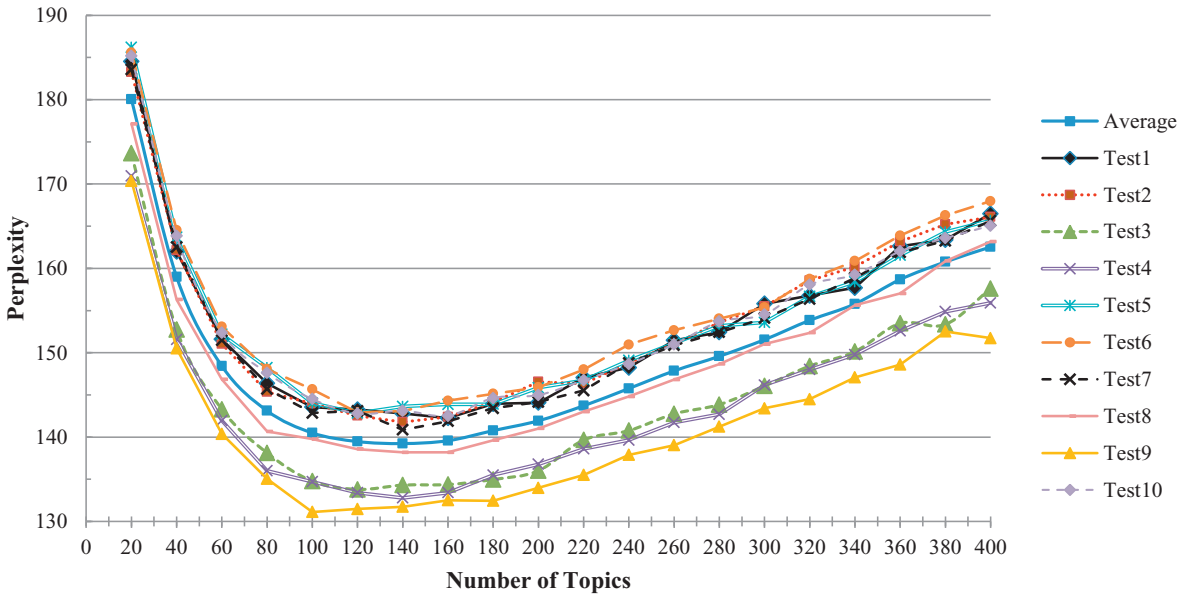

Fig. 6. Perplexity results on the public comments on B-city's official website.

Fig. 7. Interactive interface of keyword suggestion and topic selection process.

5. User evaluation experiments

To validate the usefulness and effectiveness of our proposed semantic search method based on LDA modeling (SS-LDA), we conduct a user evaluation experiment to compare with the Keywords-Matching approach (KM), which is one of the most commonly used search methods in the domain of public management and public affairs. There must to point out that KM is not the current methods used by the target sector, the office of the website engaged in related work in manual mode before this study.

id	Quarter
1	2006.5-9
2	2006.10-12
3	2007.1-3
4	2007.4-6
5	2007.7-9
6	2007.10-12
7	2008.1-3
8	2008.4-6
9	2008.7-9
10	2008.10-12
11	2009.1-3
12	2009.4-6
13	2009.7-9
14	2009.10-12
15	2010.1-3
16	2010.4-6
17	2010.7-9
18	2010.10-12
19	2011.1-3
20	2011.4-6
21	2011.7-9
22	2011.10-12
23	2012.1-3
24	2012.4-6
25	2012.7-9
26	2012.10-12
27	2013.1-4

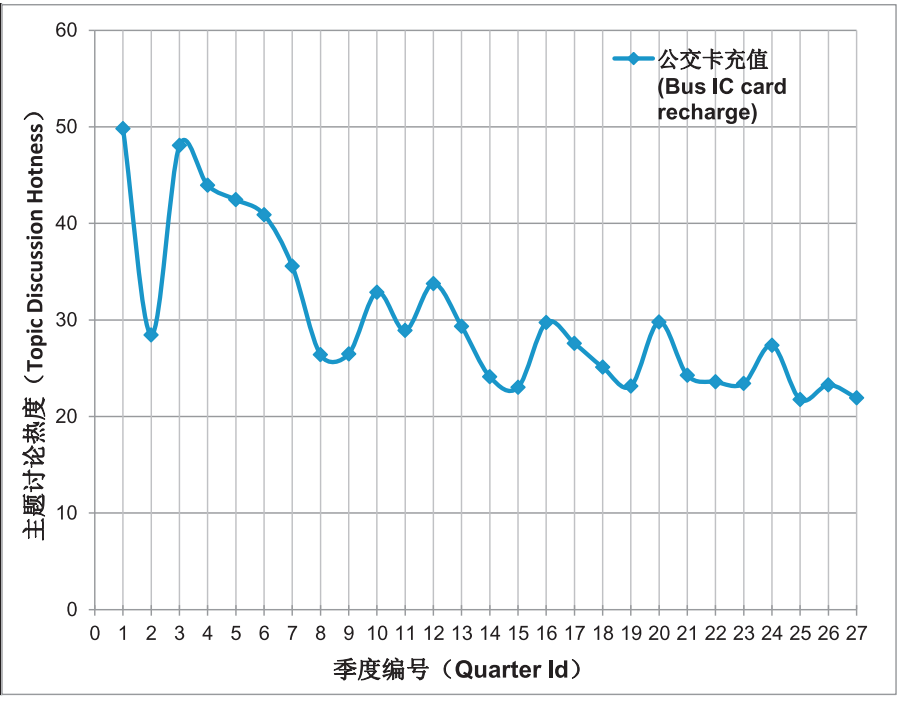

Fig. 8. Topic discussion hotness changing curve for the topic "bus IC card recharge".

用户您好！根据您已经选择的感兴趣主题，系统为您提供了如下已经排序的搜索结果，希望能够满足您的搜索需求! 谢谢您的使用！^_^
(Welcome to our system! Based on your selected topic(s) of interest, our system has provided the following list of ranked search results.
We hope it will be useful! Thank you for your participation ^_^)

ID	日期(Date)	信件标题(Comment Title)	相关性(Relevance)
1	2012/6/25	为什么本身完整的公交IC卡的换卡要如此苛刻	0.383865
2	2007/12/13	北京公交一卡通退卡业务中存在的问题	0.367733
3	2009/3/12	投诉中信银行交通一卡通充值服务	0.336648
4	2010/11/9	投诉一卡通失磁卡扣押金	0.334848
5	2007/12/13	北京公交一卡通是否可以透支	0.334656
6	2008/10/14	公交卡投诉	0.315359
7	2009/2/20	建议北京移动保留出售硬板充值卡	0.309397
8	2009/10/18	市政交通一卡通使用的不便	0.297191
9	2008/11/4	建议公交一卡通可以与牡丹交通卡合并	0.296763
10	2010/4/4	公交一卡通为什么不是实行挂失制度	0.29625

Go to page: 1 of 500 Go | Next>

返回 (Return) 重新查询 (New Search)

Fig. 9. The ranked list of search results semantically related to "bus IC card recharge" (Top 10).

Table 1
10 search tasks or public affairs topics used in the user experiment.

ID	Search task or public affairs topic
1	Bus route (公交线路)
2	Registered permanent residence or "Hukou" (户口)
3	City development (城市发展)
4	Industrial and commercial law enforcement (工商执法)
5	House purchase or renting (买房/租房)
6	Illegal occupation (违章占用)
7	Endowment or social insurance (养老社保)
8	Environmental pollution (环境污染)
9	Medical treatment and public health (公共医疗卫生)
10	Traffic jams (交通拥堵)

The performance of the contrasting KM method is based on two principles: First, any valid search results provided by the KM method should contain all of the query or search keywords; Second, the ranking of each valid search result is determined by the well-known Okapi BM25 score between the document and the query or search keywords (Robertson, Walker, Jones, Hancock-Beaulieu, & Gatford, 1994). The Okapi BM25 is a bag-of-words retrieval function (i.e., not semantic search) that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship of the query terms within a document (e.g., their relative proximity), which has been usually used as one of most effective keyword-based search models (An & Huang, 2013; Dang, Luk, & Allan, 2014; Feuer, Savev, & Aslam, 2009; Lee, Seo, Jeon, & Rim, 2011; Whissell & Clarke, 2013; Zareh Bidoki, Ghodsnia, Yazdani, & Oroumchian, 2010). Okapi BM25 is not a single function, but instead an entire family of scoring functions with slightly different components and parameters. One of the most prominent representations of the Okapi BM25 function is shown in Eq. (11).

$$BM25(Q, d_i) = \sum_{j \in Q} \frac{tf_{ij}(k_1 + 1)}{tf_{ij} + k_1\left((1 - b) + b\frac{dl_i}{avgdl}\right)} \log\left(\frac{n}{n_j}\right) \tag{11}$$

where Q is a query consisting of a set of keywords, d_i is the ith document in the document set, tf_{ij} represents the term frequency of the jth query keyword in the ith document, dl_i represents the ith document's length (i.e., number of words in the document), $avgdl$ is the average document length for documents in the collection, and n and n_j are the number of all of the documents and the number of documents containing the jth query keyword, respectively. Moreover, b and k_1 are parameters that are tuned with $k_1 \geq 0$ and $0 \leq b \leq 1$. Typical values for the BM25 parameters in document retrieval are $b = 0.75$ and $k_1 = 2.0$, and previous BM25 related papers have all used these default values (see the above BM25 papers); we adopt the same values in this user experiment.

To conduct the user experiment in the practical environment of the search system, we invited ten administrative staff members from the department responsible for B-city's online citizen opinions acquisition platform as the subjects, and divided them into two groups. The five members of the first group work on the team whose routine task is the manual classification and generalization of citizens' comments on the platform. Members of this group are considered to better understand citizens' comments (due to their own related experiences) than the other group, which consists of five staff members in the same department who conduct other tasks.

We asked each subject to perform 10 search tasks independently. Each search task is based on a typical and popular topic selected from the public affairs that are issues in B-city (see Table 1). For each search task, the subject can submit any keywords related to the corresponding topic, and the experiment system will return the top 10 search results by either our proposed SS-LDA method or by the KM method, which is essentially random and is not shown to the subjects. We need only guarantee that the numbers of the results provided by the two search methods are equal (i.e., 50 for SS-LDA and 50 for KM). Moreover, to maintain consistent result styles, the topic discussion-hotness changing curves (similar to Fig. 5) generated by SS-LDA are not provided to the subjects. In addition, to make the experiment process simple, for both search methods only the top 10 search results (i.e., the first result page) are provided.

For each search task, after the subject read the search results, he/she was asked to evaluate the results with respect to information accuracy and information richness. The two questions were measured using a seven-point Likert-type scale, ranging from "very dissatisfied" (1) to "very satisfied" (7). The accuracy in our user experiment means the results quality of being true or correct. In this dimension, the subjects should pay attention to each result shown in the system correctly match the task topic or not, and conduct an overall judgement. The richness in our user experiment means the results quality of being comprehensive or completed. In this dimension, the subjects should pay attention to all results shown in the system is enough or not for the task topic, and conduct an overall judgement. In the questionnaire, we provide the clear definitions and need explanations of the two concepts. Therefore, although only the one question for each factor, the accuracy and richness, was measured using a seven-point Likert-type scale, ranging from "very dissatisfied" (1) to "very satisfied" (7), those overall judgements could represent the subject's evaluation on the results. The subject was not asked to compare the results of two methods (LDA versus KM), but asked to compare the results he/she saw in the system to the best matching results in his/her mind. In fact, in order to avoid possible interference, the subject did not know that the judgment result is calculated with what kind of method when he/she participate in the experiment.

Table 2
Equal variance assumed test results.

Group	Method	Accuracy			Richness		
		Mean (S. D.)	EVA		Mean (S. D.)	EVA	
			t-test	*p*-value		*t*-test	*p*-value
All samples	SS-LDA	5.18 (1.58)	1.17	0.245	5.48 (1.18)	2.77	0.007
	KM	4.80 (1.65)			4.70 (1.61)		
Group1 (experienced users)	SS-LDA	4.92 (1.55)	2.27	0.028	5.12 (1.07)	3.19	0.002
	KM	3.92 (1.59)			3.92 (1.56)		
Group2 (non-experienced users)	SS-LDA	5.44 (1.61)	0.43	0.667	5.88 (1.19)	1.28	0.208
	KM	5.62 (1.27)			5.42 (1.30)		

As shown in Table 2, the user evaluations of the results of SS-LDA method are higher than those of KM method both on information accuracy and information richness. According to the equal variance assumed (EVA) test, the differences between the two methods are more significant for information richness. Another noteworthy phenomenon is the impact on the subjects' relevant experiences is that the mean scores given by the first group are higher than those given by the second group, indicating that staff who are more familiar with comments have higher demands for search tool ability.

Although the second group of users is unable to distinguish the differences between the two methods, the first group of users confirms the more significant advantages of SS-LDA, which also shows that the advantage of LDA-related methods is their in-depth analysis of the text. We put the subjects into two groups experienced and non-experienced at first because the team whose routine task is the manual classification and generalization of citizens' comments on the platform only have five people (experienced users). In order to appropriately increase the sample size, we invited to five staff members in the same department who conduct other tasks (non-experienced users) join the experiment. The experiment results show the experienced users confirm the more significant advantages of SS-LDA than non-experienced. One possible explanation is that the superiority of the LDA method may be reflected in a more professional situation. The conjecture remains to be further validation in the on-going research.

The test results show the value of the search system for the departments. This system will soon be delivered to the related departments to enhance their efficiency in dealing with citizen comments.

6. Conclusions and future work

In addressing municipal administrative departments' semantic search requirements related to the contents of an official online platform for citizens to comment on urban public affairs, this study establishes a search process framework that includes and coordinates the user search process and the probabilistic topic modeling process. This paper also describes the technical details of that framework and presents an approach of orienting the search system toward hot topics. Although such a semantic search tool is not yet routinely employed in relevant departments, the user evaluation experiment shows that potential users are more satisfied by the search results returned by SS-LDA than those returned by the KM method. The rapid search for hot topics and the analysis of variation tendency enable decision-makers to collect from the online platform information that is more complicated and that relates to policy adjustment and urban governance. However, this paper's proposed solution is only validated for a special application scenario, and a deeper discussion of its general applicability remains to be performed in subsequent work.

Broadly speaking, this paper also attempts to offer a new method for semantic search researchers that explore search solutions for specific domains or contexts (i.e., vertical search engines), rather than pursuing the widespread applicability of universal search tools (e.g., Google) in cross-media and cross-platforms. We believe that this will become a new area of growth for the technological development of semantic search because research and practices related to semantic search cannot be separated from analysis on its specific contents. Furthermore, the precision of the content analysis is obviously associated with the scenario and the context. Development of specific requirements can definitely facilitate better satisfying the end-users with the expected objective at the same technological level. We expect this paper can also provide helpful insights for subsequent, related studies.

For future work, we anticipate incorporating the dynamic topic model (Blei & Lafferty, 2006) to our system to better demonstrate evolution of topics over time. In addition, the way of choosing and visualization of thematic keywords and corresponding topics in our semantic search system can be improved by more effective measures and methods (Chuang, Manning, & Heer, 2012; Sievert & Shirley, 2014).

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China (71402007/71473143/71271044/71102010), the Beijing Social Science Foundation (15JGA008), and the Tsinghua University Initiative Scientific Research Program (20131089260). The authors would like to thank the anonymous reviewers for their thoughtful comments and suggestions.

References

Allan, J., & Raghavan, H. (2002). Using part-of-speech patterns to reduce query ambiguity. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 307–314). ACM.

An, X., & Huang, J. X. (2013). Boosting novelty for biomedical information retrieval through probabilistic latent semantic analysis. In *Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval* (pp. 829–832). ACM.

Balakrishnan, N., & Nevzorov, V. B. (2005). *A primer on statistical distributions*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Bicego, M., Lovato, P., Perina, A., Fasoli, M., Delledonne, M., Pezzotti, M., et al. (2012). Investigating topic models' capabilities in expression microarray data classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 9*, 1831–1836.

Blanco, R., Halpin, H., Herzig, D. M., Mika, P., Pound, J., Thompson, H. S., et al. (2013). Repeatable and reliable semantic search evaluation. *Web Semantics: Science, Services and Agents on the World Wide Web, 21*, 14–29.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*, 77–84.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning* (pp. 113–120). ACM.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993–1022.

Burke, M. (2009). The semantic web and the digital library. *Aslib Proceedings, 61*, 316–322.

Carterette, B., & Chandar, P. (2009). Probabilistic models of ranking novel documents for faceted topic retrieval. In *Proceeding of the 18th ACM conference on information and knowledge management* (pp. 1287–1296). ACM.

Casella, G., & George, E. I. (1992). Explaining the Gibbs Sampler. *The American Statistician, 46*, 167–174.

Chiang, R. H. L., Chua, C. E. H., & Storey, V. C. (2001). A smart web query method for semantic retrieval of web data. *Data & Knowledge Engineering, 38*, 63–84.

Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the international working conference on advanced visual interfaces* (pp. 74–77). ACM.

Dang, E. K. F., Luk, R. W. P., & Allan, J. (2014). Beyond bag-of-words: bigram-enhanced context-dependent term weights. *Journal of the Association for Information Science and Technology, 65*, 1134–1148.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*, 391–407.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*, 1–38.

Ding, L., Kolari, P., Ding, Z., & Avancha, S. (2007). Using Ontologies in the semantic web: a survey. In R. Sharman, R. Kishore, & R. Ramesh (Eds.), *Ontologies: Vol. 14* (pp. 79–113). US: Springer.

Dong, H., Hussain, F. K., & Chang, E. (2008). A survey in semantic search technologies. In *The Second IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2008)* (pp. 403–408).

Du, W., Tan, S., Cheng, X., & Yun, X. (2010). Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. In *Proceedings of the third ACM international conference on web search and data mining* (pp. 111–120). ACM.

Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., & Motta, E. (2011). Semantically enhanced information retrieval: an ontology-based approach. *Web Semantics: Science, Services and Agents on the World Wide Web, 9*, 434–452.

Feuer, A., Savev, S., & Aslam, J. A. (2009). Implementing and evaluating phrasal query suggestions for proximity search. *Information Systems, 34*, 711–723.

Fine, S., Singer, Y., & Tishby, N. (1998). The hierarchical hidden markov model: analysis and applications. *Machine Learning, 32*, 41–62.

Gao, R., Hao, B., Bai, S., Li, L., Li, A., & Zhu, T. (2013). Improving user profile with personality traits predicted from social media content. In *Proceedings of the 7th ACM conference on recommender systems* (pp. 355–358). ACM.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. In *Proceedings of the national academy of sciences of the United States of America: Vol. 101* (pp. 5228–5235).

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.). New York, NY: Springer-Verlag.

Hofmann, T. (1999). Probabilistic latent semantic analysis. *Proceedings of the fifteenth conference on Uncertainty in artificial intelligence* (pp. 289–296). Stockholm, Sweden: Morgan Kaufmann Publishers Inc.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 50–57). ACM.

Hong, H. (2013). Government websites and social media's influence on government-public relationships. *Public Relations Review, 39*, 346–356.

Horton, S. (2006). Social capital, government policy and public value: implications for archive service delivery. *Aslib Proceedings, 58*, 502–512.

Ingawale, M., Dutta, A., Roy, R., & Seetharaman, P. (2013). Network analysis of user generated content quality in Wikipedia. *Online Information Review, 37*, 602–619.

Jiang, X., & Tan, A.-H. (2009). Learning and inferencing in user ontology for personalized semantic web search. *Information Sciences, 179*, 2794–2808.

Johnson Lim, S. C., Liu, Y., & Lee, W. B. (2010). Multi-facet product information search and retrieval using semantically annotated product family ontology. *Information Processing & Management, 46*, 479–493.

Kara, S., Alan, Ö., Sabuncu, O., Akpınar, S., Cicekli, N. K., & Alpaslan, F. N. (2012). An ontology-based retrieval system using semantic indexing. *Information Systems, 37*, 294–305.

Kim, H. H. (2005). ONTOWEB: implementing an ontology-based web retrieval system. *Journal of the American Society for Information Science and Technology, 56*, 1167–1176.

La Rosa, M., Fiannaca, A., Rizzo, R., & Urso, A. (2014). Genomic sequence classification using probabilistic topic modeling. *Computational intelligence methods for bioinformatics and biostatistics* (pp. 49–61). Springer.

Lee, J.-T., Seo, J., Jeon, J., & Rim, H.-C. (2011). Sentence-based relevance flow analysis for high accuracy retrieval. *Journal of the American Society for Information Science and Technology, 62*, 1666–1675.

Lee, J., Min, J.-K., Oh, A., & Chung, C.-W. (2014). Effective ranking and search techniques for web resources considering semantic relationships. *Information Processing & Management, 50*, 132–155.

Levy, K. E. C., & Franklin, M. (2014). Driving regulation: using topic models to examine political contention in the U.S. trucking industry. *Social Science Computer Review, 32*, 182–194.

Li, D., He, B., Ding, Y., Tang, J., Sugimoto, C., Qin, Z., Yan, E., Li, J., & Dong, T. (2010). Community-based topic modeling for social tagging. In *Proceedings of the 19th ACM international conference on information and knowledge management* (pp. 1565–1568). ACM.

Li, D., Lv, Q., Xie, X., Shang, L., Xia, H., Lu, T., & Gu, N. (2012). Interest-based real-time content recommendation in online social communities. *Knowledge-Based Systems, 28*, 1–12.

Li, P.-F., Zhu, Q.-M., & Zhou, G.-D. (2014). Using compositional semantics and discourse consistency to improve chinese trigger identification. *Information Processing & Management, 50*, 399–415.

Li, Y., Ma, S., Zhang, Y., Huang, R., & Kinshuk (2013). An improved mix framework for opinion leader identification in online learning communities. *Knowledge-Based Systems, 43*, 43–51.

Linders, D. (2012). From e-government to we-government: defining a typology for citizen coproduction in the age of social media. *Government Information Quarterly, 29*, 446–454.

Linstead, E., Lopes, C., & Baldi, P. (2008). An Application of latent Dirichlet allocation to analyzing software evolution. In *Proceedings of the seventh international conference on machine learning and applications, 2008. ICMLA '08* (pp. 813–818).

Liu, B. (2007). *Web data mining: exploring hyperlinks, contents, and usage data*. BerlinHeidelberg: Springer.

Liu, B., Liu, L., Tsykin, A., Goodall, G. J., Green, J. E., Zhu, M., et al. (2010). Identifying functional miRNA–mRNA regulatory modules with correspondence latent dirichlet allocation. *Bioinformatics, 26*, 3105–3111.

Liu, H., He, J., Wang, T., Song, W., & Du, X. (2013). Combining user preferences and user opinions for accurate recommendation. *Electronic Commerce Research and Applications, 12*, 14–23.
Liu, X., & Turtle, H. (2013). Real-time user interest modeling for real-time ranking. *Journal of the American Society for Information Science and Technology, 64*, 1557–1576.
Mäkelä, E. (2005). Survey of semantic search research. In *Proceedings of the seminar on knowledge management on the semantic web: department of computer science*. University of Helsinki.
Mandala, R., Takenobu, T., & Hozumi, T. (1998). The use of wordnet in information retrieval. In *Proceedings of the conference on use of wordnet in natural language processing systems* (pp. 31–37).
Mangold, C. (2007). A survey and classification of semantic search approaches. *International Journal of Metadata, Semantics and Ontologies, 2*, 23–34.
Misra, H., Yvon, F., Cappé, O., & Jose, J. (2011). Text segmentation: A topic modeling perspective. *Information Processing & Management, 47*, 528–544.
Olszewski, D. (2012). A probabilistic approach to fraud detection in telecommunications. *Knowledge-Based Systems, 26*, 246–258.
Papadimitriou, C. H., Tamaki, H., Raghavan, P., & Vempala, S. (1998). Latent semantic indexing: a probabilistic analysis. In *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems* (pp. 159–168). ACM.
Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on world wide web* (pp. 91–100). ACM.
Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008). Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 569–577). ACM.
Ramage, D., Rosen, E., Chuang, J., Manning, C. D., & McFarland, D. A. (2009). Topic modeling for the social sciences. In *NIPS 2009 workshop on applications for topic models: text and beyond: Vol. 5*.
Rao, Y., Li, Q., Mao, X., & Wenyin, L. (2014). Sentiment topic models for social emotion mining. *Information Sciences, 266*, 90–100.
Rhoda, C. J., & Norman, A. J. (2013). Big data and transformational government. *IT Professional Magazine, 15*, 43–48.
Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science, 27*, 129–146.
Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., & Gatford, M. (1994). Okapi at TREC-3. In *The third text retrieval conference (TREC '94)*.
Ruiz-Martínez, J. M., Valencia-García, R., Martínez-Béjar, R., & Hoffmann, A. (2012). BioOntoVerb: a top level ontology based framework to populate biomedical ontologies from texts. *Knowledge-Based Systems, 36*, 68–80.
Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communication of the ACM, 18*, 613–620.
Shapiro, R. B., & Ossorio, P. N. (2013). Regulation of online social network studies. *Science, 339*, 144–145.
Shi, L., & Nie, J.-Y. (2009). Integrating phrase inseparability in phrase-based model. In *Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval* (pp. 708–709). ACM.
Sievert, C., & Shirley, K. E. (2014). LDAvis: a method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63–70).
Somasundaram, K., & Murphy, G. C. (2012). Automatic categorization of bug reports using latent Dirichlet allocation. In *Proceedings of the 5th India software engineering conference* (pp. 125–130). ACM.
Thomas, S. W., Adams, B., Hassan, A. E., & Blostein, D. (2014). Studying software evolution using topic models. *Science of Computer Programming, 80*, 457–479 *Part B*.
Toutanova, K., & Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the joint SIGDAT conference on empirical methods in natural language processing and very large corpora (EMNLP/VLC-2000)* (pp. 63–70).
Urbain, J., Goharian, N., & Frieder, O. (2008). Probabilistic passage models for semantic search of genomics literature. *Journal of the American Society for Information Science and Technology, 59*, 2008–2023.
Vandic, D., van Dam, J.-W., & Frasincar, F. (2012). Faceted product search powered by the Semantic Web. *Decision Support Systems, 53*, 425–437.
Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 448–456). ACM.
Wei, X., & Croft, W. B. (2006). LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 178–185). ACM.
Whissell, J. S., & Clarke, C. L. A. (2013). Effective measures for inter-document similarity. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management* (pp. 1361–1370). ACM.
Williams, R., Wiele, T. v. d., Iwaarden, J. v., & Eldridge, S. (2010). The importance of user-generated content: the case of hotels. *The TQM Journal, 22*, 117–128.
Wu, R.-S., & Chou, P.-H. (2011). Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. *Electronic Commerce Research and Applications, 10*, 331–341.
Xing, D., & Girolami, M. (2007). Employing Latent Dirichlet Allocation for fraud detection in telecommunications. *Pattern Recognition Letters, 28*, 1727–1734.
Xu, H., Zhang, F., & Wang, W. (2015). Implicit feature identification in Chinese reviews using explicit topic mining model. *Knowledge-Based Systems, 76*, 166–175.
Yu-Che, C., & Tsui-Chuan, H. (2014). Big Data for digital government: opportunities, challenges, and strategies. *International Journal of Public Administration in the Digital Age, 1*, 1–14.
Zareh Bidoki, A. M., Ghodsnia, P., Yazdani, N., & Oroumchian, F. (2010). A3CRank: an adaptive ranking method based on connectivity, content and click-through data. *Information Processing & Management, 46*, 159–169.
Zhang, C., & Sun, J. (2012). Large scale microblog mining using distributed MB-LDA. In *Proceedings of the 21st international conference companion on World Wide Web* (pp. 1035–1042). ACM.
Zhang, H., Yu, H., Xiong, D., & Liu, Q. (2003). HHMM-based Chinese lexical analyzer ICTCLAS. In *Proceedings of the second SIGHAN workshop on Chinese language processing* (pp. 184–187). Association for Computational Linguistics.
Zheng, H.-T., Chen, J.-Y., & Jiang, Y. (2012). An ontology-based approach to Chinese semantic advertising. *Information Sciences, 216*, 138–154.
Zhu, J. J. H., Mo, Qian, Wang, Fang, & Lu, Heng (2011). A random digit search (rds) method for sampling of blogs and other user-generated content. *Social Science Computer Review, 29*, 327–339.