Extracting Topic-Sensitive Content from Textual Documents

- A Hybrid Topic Model Approach

Yan Liang[1], Ying Liu[1*], Chong Chen[1], Zhigang Jiang[2]

[1] Institute of Mechanical and Manufacturing Engineering
School of Engineering
Cardiff University
Cardiff CF24 3AA, UK
Email: LiuY81@cardiff.ac.uk
Tel: +44-(0)29-20874696

[2] College of Machinery and Automation
Wuhan University of Science & Technology
Wuhan 430081, China

Abstract

When exploring information of a topic, users often concern its different aspects. For instance, product designers are interested in seeking information of specific topic aspects such as technical challenge and usability from online consumer opinions, while potential buyers wish to obtain general sentiment of public opinions. In this paper, we study an interesting problem called topic-sensitive content extraction (TSCE). TSCE aims to extract contents that are relevant to the samples of topic aspects highlighted by users from a single document in a given text collection. To tackle TSCE, we have proposed a new hybrid topic model which integrates different structures in both topic space and context space. It focuses on identifying contents associated with a specified topic aspect from each document. By modeling gradient documents via term profiles for context modeling and by leveraging local and global differences between probability distributions over words in both topic modeling and context modeling, it has better captured the features of various language patterns. Hence, sentence relevance ranking according to a specific topic aspect is largely improved. The experimental studies on extracting critical contents of specific aspects, including motivation and design solution, from technical patents for design analysis have shown the merits of the proposed modeling.

Keywords:

Topic-sensitive content, probabilistic topic modeling, topic network, semi-supervised

[*] Corresponding author

Notations

z_k	Latent topic variable $z_k \in \{z_1, \ldots, z_K\}$
w_j	A word $w_j \in \{w_1, \ldots, w_M\}$ in a vocabulary with M words
d_i	Document $d_i \in \{d_1, \ldots, d_N\}$
$n(d_i, w_j)$	Occurrences of word w_j in document d_i
$P(w\|z)$	Topic word distributions
$P(z\|d)$	Document topic distributions

1. Introduction

With the wide use of information technology and the great advancement of WWW and its application, information processing and filtering from a large amount of documents that often exist in a digital form has become a great challenge. These documents range from online newspapers and social nets to technical patents and academic reports. The sheer size of such document collections and their frequent pace in content update have made it more difficult for users to wade through in seeking their information of interest. Many efforts have been invested in facilitating users in this regard by discovering and extracting meaningful information, such as retrieving information at document level, multi-document summarization, extracting relevant passages at segment level, extracting entities and relations at semantic level, and topic modeling. While current techniques and approaches have helped users exploit a large number of documents at different levels, we have observed a rising interest in identifying topic-centric content.

In reality, users may have different aspects of concern when exploring a topic. Such topical aspects can be very general or more specific. A general aspect refers to a common aspect of a topic, which is relatively known to the public, such as product features and hotel facilities mentioned in online review data. For example, customers often wish to find out the public opinions, e.g., positive, negative and neutral, with respect to different product features, e.g., camera lens and screen resolution. In this context, studies like constructing an aspect-dependent sentiment lexicon for sentiment analysis applications (Lu, Castellanos, Dayal, & Zhai, 2011) and discovery of user ratings for product aspects (H. Wang, Lu, & Zhai, 2010), have been carried out. In the meantime, specific aspects of a topic refer to those more detailed and sometimes essential subtopics, such as reasons of purchasing, design issues, technical aspects are often concerned by professionals. For example, from design point of view, designers intend to understand the reasons behind certain opinions, i.e., why customers like or dislike certain features. Such an exploration can provide insights to understand customers' concerns, and better help to analyze their needs and preferences towards product design and development, and marketing and sales also.

In this paper, we report our attempt to support users in searching information related to certain topical aspects that users feel interested in. We propose to study this problem which targets at the modeling of language patterns associated with topical aspects. We demonstrate its application in the generic context of targeted content extraction and retrieval. In our study, users are allowed to supply sample segments as the instances of a topical aspect that they feel interested. By extracting contents closely relevant to the topical aspect of interest interactively, it saves much time and effort in identifying and providing targeted aspect information. To achieve this, we focus on a different text mining issue, called topical-sensitive content extraction (TSCE). Given a set of documents for an entity or a domain (e.g., MP3 players or printer) as well as some segments as examples of a topical aspect provided by users, TSCE aims at extracting contents from each single document in the collection based on how closely these contents are related to the specific topical aspect.

Revealing contents associated with various topical aspects in documents would offer a considerable advantage. One perspective is that such topic-sensitive contents extracted can serve as a summary derived from the text but are tailored more towards a specific topic aspect. It helps users to gain more focused information compared to a standard single document summary. Examples similar to this scenario are many. For instance, for prior art search in engineering design, the content of a specific design document can be written from different perspectives, such as those of motivation, the design argument and technical solution. The detailed and focused contents of motivation aspect can help junior engineers understand why certain design issues have received more attention than others. Meanwhile, contents centered on design argument aspects help to reveal more details on the trade-off considerations of different design proposals.

While some existing studies on information extraction and topic modeling are relevant to TSCE problem to a certain degree, little work has been done attempting to extract textual contents of a topical aspect as indicated by user and to model the topic involving the aspects and context where the topic appears. To tackle TSCE, we have proposed a three-stage approach based on a new hybrid topic model with biased topic network to rank sentences in each individual document. A gradient document generation approach is first proposed based on term profiles in neighboring regions for context modeling. Secondly, by exploiting the sample segments indicated by the user as the instances of a topical aspect concerned, we propose a generic hybrid topic approach to model both topics and their contexts in documents. Finally, sentences in each single text are ranked based on the topic model and the context model derived in the second stage for topical-sensitive content extraction.

The basic idea behind our hybrid topic model is that we believe that the degree of association between a sentence and a topical aspect is not only determined by how likely the sentence is related to the topic, but also dependent upon how closely the sentence is relevant to the topical aspect of interest indicated by the user. Information about an aspect can often be revealed from context. For example, in an online camera review, “The screen is not good.” and “The screen is too small.” are two sentences about the topic “screen”. The contextual information “not good” suggests that the first sentence is more likely related to quality aspect, while “too small” gives more hints on aspects like user experience or dimension. In relation to this example, more specifically, different from existing studies on probability latent semantic analysis (PLSA) (Hofmann, 1999) and latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003) which aim at representing documents properly using topic distribution, we exploit further and distinguish topical contents in single individual document based on the distribution of topic words and the context that various aspects exist. Our proposed hybrid topic model utilizes both local and global structure of document space. In this hybrid topic model, topic modeling and context modeling are proposed to be locally biased to the language model of a specific topical aspect, while all topic models and context models should be globally different from each other.

The rest of this paper is organized as follows. In Section 2, relevant studies on extracting information of interest and topic modeling are reviewed. In Section 3, we propose and detail the three-stage framework based on our hybrid topic model newly proposed for topical-sensitive content extraction. Section 4 includes our experimental studies using design documents and results followed by discussion; and Section 5 concludes.

2. Related Work

Discovering and extracting information of interest has become more and more important as the number of text collections increases rapidly and goes beyond individual capability in processing and managing them effectively and efficiently. Many studies have been invested to assist users to locate information at different granularity levels. The research of standard document retrieval (i.e., search engines) aims at providing users with a ranked list of relevant documents (Singhal, 2001; Mavridis & Symeonidis, 2014). Some studies meet users' information needs at segment level. For example, passage retrieval looks for passages that contain pieces of information about queries (Jiang & Zhai, 2006; M. Wang & Si, 2008). Document summarization generates a concise version of text which contains the most important information selected or ranked from the original texts (Karen, 2007; Radev, Jing, Styś, & Tam, 2004). Some other research efforts provide users with sentiment-level information, such as review sentiment detection (Jindal & Liu, 2006; Tang, Tan, & Cheng, 2009) and opinion summarization (Pang & Lee, 2008; Zhan, Loh, & Liu, 2009).

Recently, as probabilistic topic modeling has received much attention, the focus is moving towards extracting information at topic level. Text classification and text clustering have been revisited using topic modeling. For example, Xue et al. (2008) attempted to study the cross-domain text classification problem by extending PLSA to integrate labeled and unlabeled data. Niu and Shi (2010) studied PLSA based on a semi-supervised algorithm for document clustering by employing the must-link supervision between two documents. In addition, the concepts of topic modeling have also been introduced in document summarization. Li et al. (2013) addressed multi-document summarization by introducing Bayesian topic models. This model makes use of sentence features, e.g., sentence position, length and sentence bigram frequency.

Some recent research work shows that leveraging topic models helps to improve key-phrase or topic extraction. In (Liu, Huang, Zheng, & Sun, 2010), Liu et al. decomposed the traditional PageRank into multiple random walks specific to various topics for keyphrase extraction. Zhao et al. (2011) also used the topical PageRank for keyphrase ranking from Twitter. Moreover, using topic modeling for opinion analysis from online reviews has also received much attention. In (H. Wang et al., 2010), Wang et al. analyzed the opinions expressed in each review at the level of topical aspects to discover ratings of various aspects as well as relative importance weights on different aspects in each review. Tang et al. (2013) exploited multiple types of contexts such as titles and users to model consensus topics from the social media data. To detect short-term cyclical topic dynamics in the user-generated content and news, Lu (2015) designed a Probit-Dirichlet hybrid allocation (PDHA) topic model which incorporates a document's temporal features.

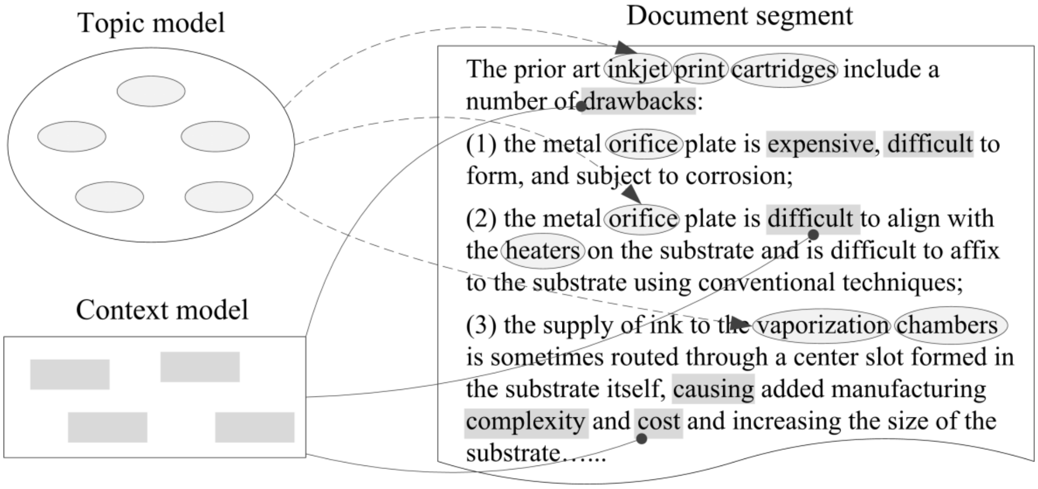
In general, topic modeling finds its way to explore the document space at topic level for document representation. In order to have superior discriminative power for document representation, several other topic models have been proposed based on PLSA. A Laplacian probabilistic latent semantic indexing (LapPLSI) is proposed for a topic modeling on the document manifold (Cai, Mei, Han, & Zhai, 2008). This method integrates the near neighbor information of a document into the log-likelihood maximization to estimate the topic model. Based on the similar idea that two sufficiently close documents should have similar topic distribution, a probabilistic dyadic data analysis approach is studied to take into account local manifold structure and global consistency (Cai, Wang, & He, 2009). This approach uses Kullback-Leibler divergence to measure the distance between two topic probability distributions in topic modeling. In addition, a locally discriminative topic model (LDTM) is proposed (Wu et al., 2012). LDTM assumes that topic distribution in an individual document strongly depends on its neighbors. It provides a complementary discriminative learning scheme to infer the topic distribution via regressions. To mine coherent topics in documents, word embeddings obtained from a large number of domains were combined into LDA approach in topic modeling (Yao et al., 2017).

We have noticed that most of the previous studies focus on discriminating documents, while few of them have attempted to recognize different segments in each single document. For our study on TSCE problems, we explore the problem on how different segments of textual contents in a single text can be modeled and differentiated. Given a limited number of text segments for the indicative texts of a topical aspect by a user, our work aims to uncover and rank text contents in each single text that look most similar to the indicative texts. In other words, they reveal the topical aspect specified by the user. The results can be useful for multiple tasks, including summarization from multiple aspects and document analysis at the topic aspect level. In our previous study, we have attempted to tackle TSCE by building semantic graphs based on a sequential langue model (Liang, Liu, Lee, & Kwong, 2011). In this paper, we further our efforts to model topics as well as contexts in a text collection for topic-sensitive content extraction.

3. A Hybrid Topic Model for Topical-Sensitive Content Extraction

3.1 Three-Stage Framework

In order to extract topic-sensitive contents, we start from analyzing the usage of terms from both topic and context perspectives. Figure 1 shows an example related to the aspect of the design issue. Topical terms, such as "inkjet" "print" and "cartridges", show that this document refers to inkjet printer design. Other terms in the contexts such as "drawbacks" "difficult" "complexity" and "cost" reveal the problem aspect of the printer design. By investigating both topical terms and their context, it can be inferred that this segment concerns about why the author intended to focus on the inkjet printer design.



The prior art inkjet print cartridges include a number of drawbacks:

(1) the metal orifice plate is expensive, difficult to form, and subject to corrosion;

(2) the metal orifice plate is difficult to align with the heaters on the substrate and is difficult to affix to the substrate using conventional techniques;

(3) the supply of ink to the vaporization chambers is sometimes routed through a center slot formed in the substrate itself, causing added manufacturing complexity and cost and increasing the size of the substrate……

Figure 1. An example of a document segment generated by topic modeling and context modeling

In this paper, we propose a hybrid topic model which makes use of both topic and context language features to extract topic-sensitive content. Our model can be considered as an extension of PLSA to explore both topic modeling and context modeling in a text collection. The basic idea is to measure how strongly a sentence can deliver the message that is closely associated with a topical aspect, where this topical aspect is described using some segments provided by the user. Figure 2 shows the three-stage framework of our approach. It includes three major steps, i.e., gradient document generation, hybrid topic modeling and sentence ranking. Firstly, a gradient document generation approach is proposed for context modeling. To suggest the features of a context, the gradient document is produced by leveraging term profiles in a neighboring region. Next, the second stage is for topic modeling and context modeling respectively using the biased topic network. Lastly, using the topic model and the derived context model, we score each sentence and rank them according to their respective sensitivity related to the topical aspect. The following sections detail our three-stage framework.

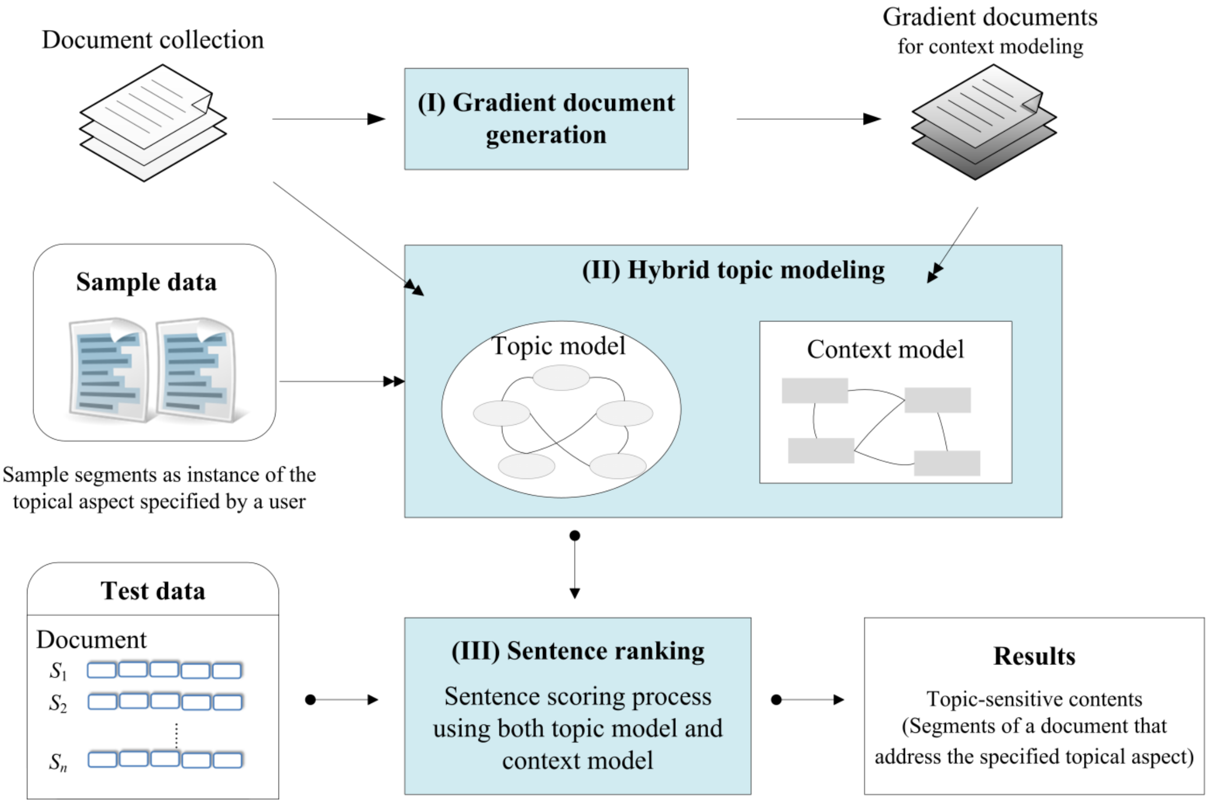

Document collection
(I) Gradient document generation
Gradient documents
for context modeling
Sample data
Sample segments as instance of the topical aspect specified by a user
(II) Hybrid topic modeling
Topic model
Context model
Test data
Document
S_1
S_2
S_n
(III) Sentence ranking
Sentence scoring process using both topic model and context model
Results
Topic-sensitive contents
(Segments of a document that address the specified topical aspect)

Figure 2. The hybrid topic modeling for topic-sensitive content extraction

3.2 Gradient Document Generation for Context Modeling

The gradient document generation step is to characterize salient features of contexts in each individual document. A context in our study refers to a linguistic context or verbal context which points to the local surrounding text of a linguistic unit that is useful to infer the meaning of a linguistic unit (i.e., a word, a concept or an entity). While the context terms are often with low term frequency, such as the example presented in Section 3.1, they sometimes are the signal of indicating a topical aspect. Inspired by edge detection in image processing which often leverages the region differences of gradient images, this step is designed by considering the difference between term frequencies in the neighboring regions. This sort of local neighborhood information is often neglected in previous topic models. In our work, aiming to recognize contents at the topic aspect level from each individual document, it leads to the idea that it would be more helpful for TSCE issue if the topic word distributions $P(w|z)$ among different topics are highly discriminative. We have noticed that term frequency $n(d_i, w_j)$ will largely affect the estimation of $P(w|z)$. Therefore, we intend to build a gradient document for context modeling.

Specifically, we assume that the ability of a term in reflecting the meaning of a topical aspect depends not only on its own frequency, but also on its neighboring terms' frequency features. Given a document d_i with a word sequence $\{t_1,\ldots, t_q, \ldots t_{|di|}\}$ shown in Figure 3, the gradient document of d_i is denoted as $n'(d_i, w_j)$ for each term w_j in d_i.

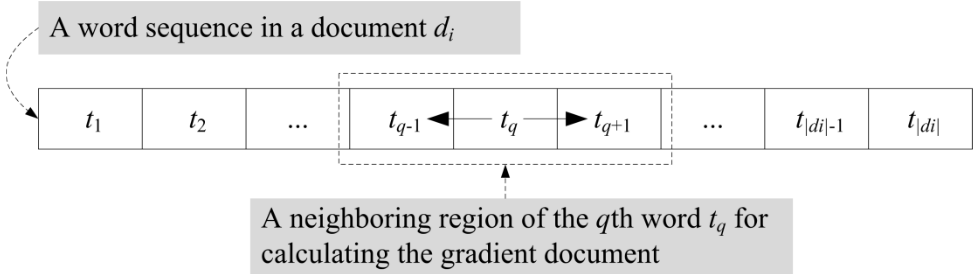

A word sequence in a document d_i
t_1
t_2
...
t_{q-1}
t_q
t_{q+1}
...
$t_{|di|-1}$
$t_{|di|}$
A neighboring region of the qth word t_q for calculating the gradient document

Figure 3. Two-directional gradient between term frequencies

We define two approaches for calculating $n'(d_i, w_j)$, i.e., $n_1'(d_i, w_j)$ as shown in Equation (1) and $n_2'(d_i, w_j)$ as shown in Equation (2).

$$n_1'(d_i, w_j) = \frac{1}{2n(d_i, w_j)} \sum_{t_q = w_j} | n(d_i, t_{q-1}) - n(d_i, t_q) | + | n(d_i, t_q) - n(d_i, t_{q+1}) | \quad (1)$$

$$n_2'(d_i,w_j)=\frac{1}{2n(d_i,w_j)}\sum_{t_q=w_j}\sqrt{|\,(n(d_i,t_{q-1})-m(t_q))^2-(n(d_i,t_{q+1})-m(t_q))^2\,|}$$
$$m(t_q)=\frac{1}{3}\left(n(d_i,t_{q-1})+n(d_i,t_q)+n(d_i,t_{q+1})\right) \tag{2}$$

The notion n_1'(d_i, w_j) is defined based on the average difference of term frequencies between term w_j and its neighboring terms. n(d_i, w_j) is the number of term w_j in document d_i. The notion n_2'(d_i, w_j) is defined based on the standard deviation of term frequencies in certain neighboring region where term w_j occurs. m(t_q) denotes the mean frequency of terms in the neighboring region of term t_q at position q. By using local term features, the gradient document n'(d_i, w_j) is envisioned to be more capable in revealing the features of a context.

3.3 Hybrid Topic Modeling and Context Modeling with Biased Topic Network

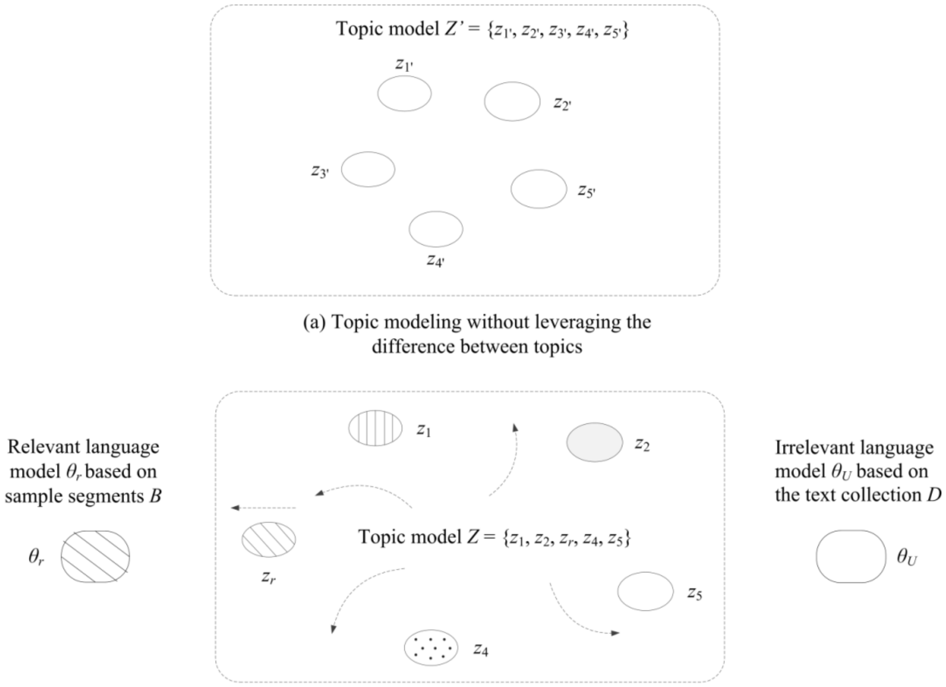
In the second stage, we propose a hybrid topic modeling approach for TSCE problem from both topic and context based on PLSA. In the topic modeling, we assume that co-occurrence data in d_i is associated with an unobserved topic variable $z_k \in Z$ (Z={z_1, …, z_K}), and similarly in context modeling, the gradient document of d_i is associated with another unobserved context variable c$_l$ ∈C (C={c_1, …, c_T}). Probability distributions over the words in both topic space and context space are analyzed. This highlights one main contribution different from the previous topic modeling approaches which mainly attempted to uncover the discriminative structure in document space.

3.3.1 Latent variable model with biased topic network

Recall that the primary task is to extract the topic-sensitive contents which address the topical aspect of interest specified. Given a domain text collection D and a sample segment set B (instances of the topical aspect specified by a user), two language models, i.e., the relevant language model θ_r based on B and the irrelevant language model θ_U based on D, are first defined to model the language patterns embedded in the specified topical aspect, shown in Equation (3). That is to say, terms which appear in the contents related to the specified topical aspect are likely to be suggested by the relevant language model θ_r, and on the contrary, they are unlikely to be suggested by the irrelevant language model θ_U.

$$p\left(w_j \middle| \theta_r\right) = \frac{n\left(B, w_j\right)}{\sum_{i=1}^{M} n\left(B, w_i\right)}, \ p\left(w_j \middle| \theta_U\right) = \frac{n\left(D, w_j\right)}{\sum_{i=1}^{M} n\left(D, w_i\right)} \quad (3)$$

Moreover, we intend to study how topic word distributions can be better exploited to estimate the patterns of term usage in different topical aspects. In proposing this hybrid topic model, we conjecture that some topic word distributions in both topic modeling and context modeling should be locally sensitive to the language models, i.e., conditional distributions in θ_r and θ_U, while simultaneously all the topic word distributions should be globally different from their homogeneous variables, as illustrated in Figure 4. In other words, for example, in topic modeling, one latent topic z_r of Z should have topic word distribution $P(w_j|z_r)$ closer to the probabilistic distribution of the sample data. In addition, if two latent topics z_{k1}, $z_{k2} \in Z$ are two different topic models, then their probability distributions over words (i.e., $P(w_j|z_{k1})$ and $P(w_j|z_{k2})$) should be different from each other. Those assumptions also hold for the context modeling we proposed.


Topic model $Z' = \{z_{1'}, z_{2'}, z_{3'}, z_{4'}, z_{5'}\}$
$z_{1'}$
$z_{2'}$
$z_{3'}$
$z_{5'}$
$z_{4'}$
Relevant language model θ_r based on sample segments B
θ_r
z_1
z_2
Topic model $Z = \{z_1, z_2, z_r, z_4, z_5\}$
z_r
z_5
z_4
Irrelevant language model θ_U based on the text collection D
θ_U

(a) Topic modeling without leveraging the difference between topics

(b) Our hybrid topic modeling considering the difference between topic word distributions

Figure 4. Topic modeling illustration. (a) Topic modeling without leveraging difference between topics. In an example of topic model *Z*' with five topics, those topics may have similar features between each other (i.e., small distance between topics as shown in (a)). (b) Our hybrid topic modeling considering difference between topic word distributions. θ_r is the relevant language model obtained based on *B* which includes segments of the topical aspect specified by a user. θ_U is the irrelevant language model which is defined based on the text collection *D*. In an example of topic model $Z = \{z_1, z_2, z_r, z_4, z_5\}$, our hybrid topic modeling would have more discriminative power between topics (i.e., large distance between topics as shown in (b)). It would also generate some topics (e.g., z_r as shown in (b)) that have a similar topic word distribution as the relevant language model θ_r for the specified topical aspect.

Then, the distance between two probability distributions over the words of two topics is defined. Given two probability distributions $P(w|z_{k1})$ and $P(w|z_{k2})$, their distance is defined in Equation (4):

$$Dis\left(P(w \mid z_{k1}), P(w \mid z_{k2})\right) = \frac{1}{2}\sum_{w_j}\left(P(w_j \mid z_{k1}) - P(w_j \mid z_{k2})\right)^2 \tag{4}$$

The assumption of regulating topic word distribution in topic modeling can be formulated by the following items in Equation (5). We first assume that there is only one topic model $z_r \in Z$ which possesses the similar probability distribution over the words as the relevant language model θ_r of the specified topical aspect. In order to simplify the estimation process, this assumption is actually presented from an opposite view, as shown in the first item on the right side of Equation (5), which is to say that topic z_r would possess a probability distribution different from the irrelevant language model θ_U. The second item suggests that except topic z_r, other topics should have a large distance with the probability distribution in the relevant language model θ_r. The third item is defined to model the difference between topics in the topic network and it sums up the distances between any two topics in topic modeling.

$$R_Z = Dis\left(P(w \mid z_r), P(w \mid \theta_U)\right) + \sum_{z_k \neq z_r} Dis\left(P(w \mid z_k), P(w \mid \theta_r)\right) + \sum_{k1,k2=1}^{K} Dis\left(P(w \mid z_{k1}), P(w \mid z_{k2})\right) \tag{5}$$

By maximizing R_Z, we intend to find out topic word distributions which hold more discriminative power between topic models and are in a better position to reveal different term’s ability in suggesting topical aspect contents. In defining our new latent variable model based on PLSA, we transfer the problem of maximizing R_Z into minimizing ($-R_Z$) and the parameter estimation for topic modeling can be solved by maximizing the log-likelihood function in Equation (6), where $\lambda > 0$:

$$\begin{aligned} Q_Z &= L_Z - \lambda(-R_Z) \\ &\propto \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j)\log\sum_{k=1}^{K} P(w_j \mid z_k)P(z_k \mid d_i) - \lambda(-R_Z) \end{aligned} \tag{6}$$

Furthermore, for context modeling, we have the following item for maximizing the difference between context topics. In Equation (7), c_r, c_l, c_{l1} and c_{l2} all belong to C, unobserved context variables.

$$R_C = Dis\left(P(w|c_r), P(w|\theta_U)\right) + \sum_{c_l \neq c_r} Dis\left(P(w|c_l), P(w|\theta_r)\right) + \sum_{l1,l2=1}^{T} Dis\left(P(w|c_{l1}), P(w|c_{l2})\right) \tag{7}$$

By transferring the maximization form of R_C into the minimization form ($-R_C$), context modeling can be formulized as maximizing the log-likelihood function shown in Equation (8), where $\lambda > 0$:

$$\begin{aligned} Q_C &= L_C - \lambda(-R_C) \\ &\propto \sum_{i=1}^{N}\sum_{j=1}^{M} n'(d_i, w_j) \log \sum_{l=1}^{T} P(w_j | c_l) P(c_l | d_i) - \lambda(-R_C) \end{aligned} \tag{8}$$

3.3.2 Model estimation with generalized EM

In our hybrid topic modeling, we need to estimate (NK+MK) parameters $\Theta = \{P(w_j|z_k), P(z_k|d_i)\}$ for topic modeling as well as (NT+MT) parameters $\Theta_C = \{P(w_j|c_l), P(c_l|d_i)\}$ for context modeling. Since we have adopted a similar approach in both model estimation processes, we present the estimation process of topic model as an example.

The standard approach for maximum likelihood estimation in latent variable model is the Expectation Maximization (EM) algorithm (Dempster et al., 1977). Given a set of initial guesses of parameters, the EM algorithm begins to update the successive estimates based on the unobserved latent variables until convergence. The EM iteration alternates between two steps: (i) an expectation (E) step where posterior probabilities for the latent variables are computed using the current estimate for the parameters, (ii) a maximization (M) step, which computes parameters by maximizing the expected complete data log-likelihood that depends on the posterior probabilities of the latent variables in the E-step.

In Equation (6), L_Z is similar with the log-likelihood function of PLSA. It represents how likely the collection of documents observed can be generated by the joint probability model $P(d_i, w_j)$. By maximizing L_Z, we can seek a set of parameters Θ satisfied. R_Z includes the discrimination between topic word distributions in local and global structures. The maximization process of $-\lambda(-R_Z)$ would lead to a better modeling of $P(w_j|z_k)$ to reveal the difference between topics.

In the following, we describe both E-step and M-step in our algorithm dedicated for parameter estimation.

E-step: compute the posterior probabilities $P(z_k|d_i, w_j)$ for the latent variables, shown in Equation (9).

$$P(z_k \mid d_i, w_j) = \frac{P(w_j \mid z_k)P(z_k \mid d_i)}{\sum_{l=1}^{K} P(w_j \mid z_l)P(z_l \mid d_i)} \tag{9}$$

M-step: maximize the expected complete data log-likelihood with constraints, shown in Equation (10)

$$\max_{\Theta} Q_Z = \max_{\Theta} L_Z - \lambda(-R_Z) \tag{10}$$

To estimate $P(z|d)$, we notice that since R_Z in our model does not involve $P(z|d)$, we can have the close form re-estimation equation for $P(z|d)$ with constraints $\sum_{k=1}^{K} P(z_k \mid d_i) = 1$ and $\sum_{j=1}^{M} P(w_j \mid z_k) = 1$. The estimation of $P(z|d)$ is the same as that of PLSA, as shown in Equation (11).

$$P(z_k \mid d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j)P(z_k \mid d_i, w_j)}{n(d_i)} \tag{11}$$

As for $P(w|z)$, we do not have the close form re-estimation equation. In this case, we cannot apply the traditional EM algorithm to estimate the parameters. Therefore, we adopt the Generalized EM (GEM) to maximize the expected complete data log-likelihood of our model in Equation (10). The major difference between GEM and traditional EM lies in the M-step. Instead of finding the global optimal solutions of *Θ* to maximize Q_Z in the M-step in traditional EM algorithm, GEM only chooses a better *Θ* that does not decrease the expected complete data log-likelihood Q_Z. Let Θ_n denote the parameters of the *n*th iteration and Θ_{n+1} denote the parameter values of the successive iteration. In GEM, the M-step chooses Θ_{n+1} which satisfies $Q_Z(\Theta_{n+1}) \geq Q_Z(\Theta_n)$ (McLachlan & Krishnan, 2008).

In estimating *P*(*w*|*z*), we start with $\Theta_{n+1}^{(1)}$ which maximizes L_Z rather of the whole Q_Z. This can be obtained by applying Equation (12) which is identical as in PLSA. Obviously, it is not guaranteed that $Q_Z(\Theta_{n+1}^{(1)}) \geq Q_Z(\Theta_n)$ holds.

$$P(w_j \mid z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j) P(z_k \mid d_i, w_j)}{\sum_{m=1}^{M} \sum_{i=1}^{N} n(d_i, w_m) P(z_k \mid d_i, w_m)} \tag{12}$$

Next, it begins with $\Theta_{n+1}^{(1)}$ and tries to decrease ($-R_Z$) by using Newton-Raphson method iteratively. Notice that R_Z only involves parameters *P*(*w*|*z*), therefore, only $P(w|z)_{n+1}$ parts in Θ_{n+1} need to be updated. Given a function *f*(*x*) and an initial x_t, the Newton-Raphson method aims to find successively better approximations to the root of the function, as shown in Equation (13), where $0 \leq \gamma \leq 1$ is the step factor.

$$x_{t+1} = x_t - \gamma \frac{f'(x)}{f''(x)} \tag{13}$$

In our case, we have $f(x) = (-R_Z)$ and $x_t = P(w_j|z_k)^t$. Based on Equation (13), we can obtain the close form solution for $P(w_j|z_k)_{n+1}^{(t+1)}$ at the (*t*+1)th iteration.

$$P(w_j \mid z_k)_{n+1}^{(t+1)} = \begin{cases} (1-\gamma) P(w_j \mid z_k)_{n+1}^{(t)} + \gamma \dfrac{P(w_j \mid \theta_U) + \sum_{l=1}^{K} P(w_j \mid z_l)_{n+1}^{(t)}}{(K+1)}, z_k = z_r \\ (1-\gamma) P(w_j \mid z_k)_{n+1}^{(t)} + \gamma \dfrac{P(w_j \mid \theta_r) + \sum_{l=1}^{K} P(w_j \mid z_l)_{n+1}^{(t)}}{(K+1)}, z_k \neq z_r \end{cases} \tag{14}$$

Clearly, both $\sum_{j=1}^{M} P(w_j \mid z_k)_{n+1}^{(t+1)} = 1$ and $P(w_j \mid z_k)_{n+1}^{(t+1)} \geq 0$ hold in Equation (14) as well as $\sum_{j=1}^{M} P(w_j \mid z_k)_{n+1}^{(t)} = 1$ and $P(w_j \mid z_k)_{n+1}^{(t)} \geq 0$. We iteratively alternate $P(w_j|z_k)$ until $Q_Z(\Theta_{n+1}^{(t+1)}) \leq Q_Z(\Theta_{n+1}^{(t)})$. Then we test whether $Q_Z(\Theta_{n+1}^{(t)}) \geq Q_Z(\Theta_n)$. If not, we discard the estimation of $\Theta_{n+1}^{(t)}$, and use Θ_n as the result of M-step. By using the values of $P(w_j|z_k)$ and $P(z_k|d_i)$ obtained in this M-step, we then continue with the next E-step. E-step and M-step are processed alternately to update parameters until convergence.

As for context modeling, the gradient document for each individual text d_i in D is first built. Then based on the gradient document $n'(d, w)$, the same parameter estimation process as topic modeling is adopted with c_l as the latent context variable and Θ_C as the parameter set.

3.4 Sentence ranking

After we obtain the hybrid topic model, each sentence in d_i needs to be scored to reflect how closely this sentence is related to the topical aspect indicated by the user's information of interest. Given sentences $S = \{s_1, \ldots, s_q, \ldots, s_{|S|}\}$ of d_i, we propose two scoring functions, i.e., $F_{main}(s_q, d_i)$ and $F_{train}(s_q, d_i)$, for sentence ranking using both topic model and context model.

The $F_{main}(s_q, d_i)$ calculates the score of a sentence s_q by using the topics that have the highest topic distribution of document d_i in topic modeling and context modeling respectively. As shown in Equation (15), we first figure out topics z_{k_m} and c_{l_m} that have the highest value of $P(z|d)$ and $P(c|d)$ respectively. By assuming that this two-topic combination can better help to recognize the topic-sensitive contents than any other topic combinations do, we score each sentence s_q by integrating all the probabilities of terms in s_q of the topic model z_{k_m} and the context model c_{l_m}, where $0 \leq \beta \leq 1$.

$$\begin{aligned} F_{main}(s_q, d_i) &= \beta \sum_{w_j \in s_q} P(w_j \mid z_{k\_m}) + (1-\beta) \sum_{w_j \in s_q} P(w_j \mid c_{l\_m}) \\ z_{k\_m} &= \arg\max_{z_k} P(z_k \mid d_i),\ c_{l\_m} = \arg\max_{c_l} P(c_l \mid d_i) \end{aligned} \tag{15}$$

Another sentence ranking approach $F_{train}(s_q, d_i)$ is designed based on the combination of a topic model and a context model that can have the highest probability in suggesting the sample segment set B specified by the user. As shown in Equation (16), we first find out the combination of topic z_{k_t} and context c_{l_t} that respectively harvests the highest summation of term probabilities of all terms in sample data B. Next, we consider that this topic combination can better help to generate the topic-sensitive contents. Hence, we score each sentence s_q based on $P(w_j|z_{k\_t})$ and $P(w_j|c_{l\_t})$. Lastly, we rank these sentences based on sentence scores for TSCE.

$$\begin{aligned}
&F_{train}(s_q, d_i) = \beta \sum_{w_j \in s_q} P(w_j \mid z_{k\_t}) + (1-\beta) \sum_{w_j \in s_q} P(w_j \mid c_{l\_t}) \\
&z_{k\_t} = \arg\max_{z_k} \sum_{w_j \in B} P(w_j \mid z_k), \; c_{l\_t} = \arg\max_{c_l} \sum_{w_j \in B} P(w_j \mid c_l)
\end{aligned} \quad (16)$$

4. Experimental Study

4.1 Experiment Setup

In our experimental study, we create a typical context in engineering design, particularly at the conceptual design stage, where designers need to do prior art search to gather detailed content information about the motivation as well as the solution-reason aspect of design from patent documents. This is because that analyzing the contents of these two topical aspects can not only help designers to understand the major issues and techniques available, but also assist designers in design knowledge reuse and design decision-making. Furthermore, it helps designers avoid any potential intellectual property conflict, and it may place the design outcome in a more favorable position for patent filing. As a matter of fact, prior art search is widely regarded as a crucial step in conceptual design, but lacks technological support.

These two topical aspects of motivation and solution-reason are chosen since they are of significant interest in conceptual design as well as in design knowledge management. In addition, in each individual patent, their relevant contents usually differ greatly in length. Hence, it helps us to analyze the merits of our approach when dealing with different lengths of contents when addressing different topical aspects.

Our dataset consists of three hundred patent documents relevant to the domain of "inkjet printer design". They were downloaded from the United States patent database (USPTO). For testing purpose, the contents of these two topical aspects in the patents were manually annotated. We have also investigated the profile of the dataset. On average, each document has 8550 words and 257 sentences. The average sentence length is 33.41 words. In addition, due to the writing style of patents, 97% of the documents contain one or more sentences with more than 100 words each. These sentences are relatively long compared to sentences in other data sources like academic journal articles. As for the motivation aspect, on average, only 3.5% of a patent document (i.e., about 250 words per document average) are tagged as motivation related. Contents related to the solution and reason aspect stand for about 15% of a document (i.e., about 1390 words per document average).

In our experiments, we treat each patent document as free text. We evaluate our approach from both performance and robustness perspectives. The performance test aims to assess how well the approach can generate results compared with human annotated contents. We use ROUGE-1 measurement which matches the unigram co-occurrences between the systems generated results and the human annotation data in terms of precision, recall and F value (Lin & Hovy, 2003). ROUGE-1 is used since it has been shown to agree with human judgment most in evaluating the machine generated text segment (Lin & Hovy, 2003). In a robustness test, we intend to evaluate the performance of an algorithm under different parameter settings. These include different number of sample data, proportion factor β, and different number of topics in our hybrid topic modeling.

4.2 Scenario I: Extract Product Design Motivations from Patents

4.2.1 Performance among relevant approaches

In the first experiment, since our task is different from other information extraction tasks and there is no report about the performance on this problem, we implement several relevant methods for comparison. The first baseline, BL-s, is a simple baseline method that takes the first *x*-word long segment from a document as the result. The next two baseline approaches are BL-TFIDF and BL-BM25. They use classic tf-idf and bm25 similarities respectively in information retrieval to rank the sentences based on the sentence similarity between the sentences and sample segments. Then we implement LDA_(F_{train}) and LDA_(F_{main}) methods based on LDA model. In LDA_(F_{train}), after obtaining the topic model using LDA, each sentence in a document is scored based on the F_{train} with $\beta = 1$ (i.e., only the topic model is used without the context model). As for LDA_(F_{main}), each sentence in the document is ranked based on F_{main} with $\beta = 1$. Another two baseline methods are PLSA_(F_{train}) and PLSA_(F_{main}). Both of them generate topic model based on PLSA, and then use F_{train} and F_{main} methods with $\beta = 1$ respectively to rank sentences of each document.

We first perform pre-processing on each document, including converting words into lower cases, removing stop words and stemming each word. Then the proposed three-stage approach starts to generate gradient document of each individual text for context modeling. Our hybrid topic modeling method is applied to obtain topic models and context models given the patent collection and sample segments of topical interest provided by subjects. In our study, human annotation data are selected as the sample segments. Lastly, for each individual document, we score sentences using the sentence ranking strategy for content extraction associated with the motivational reason aspect. Then top sentences were selected as the topic-sensitive contents until it reached *x* words, where $x = 250$. As for parameter tuning, we conduct several trials with different combinations of parameters. The parameter settings include the number of sample segments $E \in \{5, 7, 9, 12, 15\}$, topic number $K = T \in \{2, 4, 6, 8, 10, 12, 14, 16\}$ in topic modeling and context modeling, sentence scoring factor $\beta \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ and other factors $\lambda = 1$ and $\gamma = 0.1$.

For comparison purposes, Table 1 shows the best results based on the average *F* value of each approach. Our first observation is that the proposed hybrid topic model outperformed LDA and PLSA in extracting contents related to motivation aspect. The overall performance in *F* value of our approach using the scoring method F_{train} is 0.6591. It is about 13% higher than other approaches based on either LDA or PLSA. In addition, our approach obtains 1.00 in average recall and 0.4915 in average precision. These results are about 20% and 10% higher than the best results of average recall and average precision reported from LDA and PLSA respectively.

Table 1. The ROUGE-1 results for topic-sensitive content extraction associated with motivation aspect

	Average recall	Average precision	Average *F* value
Hybrid topic model_(F_{train})	**1.0000**	**0.4915**	**0.6591**
LDA_(F_{train})	0.5948	0.2887	0.3887
PLSA_(F_{train})	0.6035	0.2941	0.3955
Hybrid topic model_(F_{main})	0.8017	0.3924	0.5269
LDA_(F_{main})	0.7759	0.3782	0.5085
PLSA_(F_{main})	0.8017	0.3908	0.5254
BL-BM25	0.6638	0.3080	0.4208
BL-TFIDF	0.6466	0.3000	0.4099
BL-s	0.4435	0.2561	0.3170

It has also witnessed that our approaches, either using F_{train} or F_{main}, can extract better results. It reveals that the proposed hybrid topic model using gradient document for context modeling can better model and differentiate between topics in a text collection. Moreover, when using F_{train} for sentence scoring, we find that the hybrid topic model reaches about 25% higher in F value than LDA's 0.3887 and PLSA's 0.3955. It indicates that our approach has made good use of the sample data in language pattern learning for the specified topical aspect. Table 2 gives some example results of motivation contents extracted by our approach.

Table 2. Topic combination for motivation aspect content extraction and some examples of motivation aspect contents extracted by our hybrid topic model_(F_{train})

Topic combination for motivation aspect content extraction (The 25 most representative terms generated by our hybrid topic model_(F_{train}) for topic combination from topic model and the context model respectively. The terms are selected according to the probability $P(w\|z)$ and $P(w\|c)$.)	
z_{k_t} in topic model	**c_{l_t} in context model**
ink, print, nozzl, head, pressur, liquid, time, record, form, suppli, chamber, oper, invent, embodi, cartridg, step, control, eject, unit, surfac, printer, posit, portion, passag, amount	ink, suppli, plural, flow, form, includ, surfac, compris, droplet, nozzl, chamber, jet, background, us, print, contain, correspond, printhead, eject, commun, respect, amount, inkjet, separ, caus
Motivation aspect content extracted from patent documents (Some representative segments from the extracted results.)	
ID 1	This invention relates to thermal inkjet printing and, more particularly, to detecting the sufficiency of ink flow through the printhead of a thermal printing device such as a computer printer, facsimile machine or the like. An object of the invention, therefor, is to provide a reliable method of detecting the sufficiency of ink flow through a thermal inkjet printhead which overcomes the drawbacks of the prior art.
ID 20	Furthermore, there is a problem in that the following operation cannot be readily performed due to its configuration: a wiping operation for removing paper or ink dust, which causes nozzle plugging from the ejection face with a piece of rubber or felt.
ID 22	It is an object of the present invention to provide a method of manufacturing a dielectric film capable of controlling a crystalline state relatively easily and thereby obtaining stable characteristics constantly, and a method of manufacturing a liquid jet head capable of enhancing characteristics of a piezoelectric element.

4.2.2 Performance under different lengths of contents

In the second experiment, we compare the performance of our approaches with other relevant methods shown in Table 1 by selecting different length of segments to form topic-sensitive contents. The choices of different segment length x allows us to investigate the performance of an approach in terms of its effectiveness, i.e., whether the desired contents can be secured in the higher rank of the results.

Figure 5 shows the average F values of our approaches and other relevant methods when different lengths of segments are selected as the topic-sensitive contents for each individual document. The first observation is that when the segment length is $x = 100$, our approach can secure contents that are more relevant to the human annotation in the higher position of the ranked sequence of sentences. It obtains about 0.55 in average F value which is about 15% higher than the results generated by either LDA or PLSA. Even as more words are included ($x > 100$), our approaches using F_{train} and F_{main} are able to hold the first place among different approaches. In addition, we notice that the hybrid topic model can deliver better results when using F_{train} than F_{main}. It indicates that in our approach, a more reasonable topic combination can be chosen by leveraging sample data as shown in F_{train}, rather than just simply selecting the main topic based on the highest document topic probability shown in F_{main}.

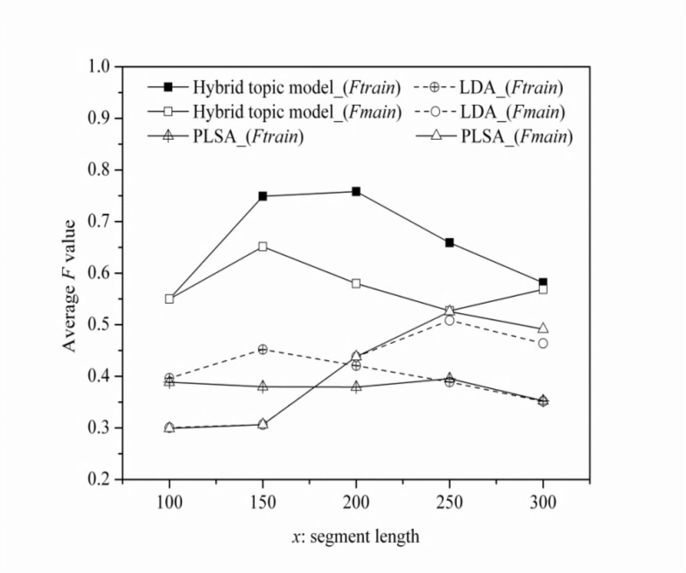
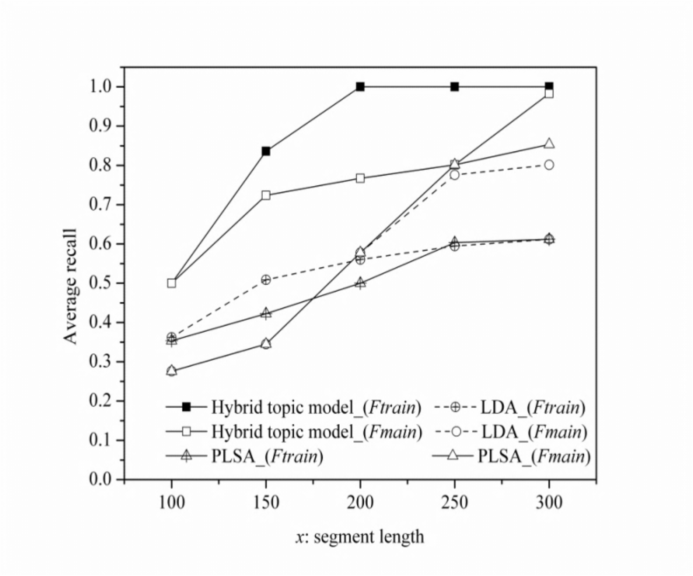


Figure 5. The average *F* values with different lengths of segments selected for motivation content extraction

Figure 6 shows the average recall and precision values with respect to different lengths of segments selected. We observe that in the same lengths of segments, our approaches can generate both higher precision and recall performance compared with other relevant approaches. In terms of average recall, our approach based on F_{train} can continue select positive terms as *x* increases from 100 to 300. In addition, in terms of average precision, when *x* increases to 150, it can secure contents with about 70% positive matched with the human annotation data. It is about 30% higher than the best results generated by LDA and PLSA. Although the average precision of our approaches decreases as *x* increases from 150 to 250, they still outperform others. It reveals that the proposed approach can identify topic-sensitive contents at the earlier stage of the ranking results when compared to other approaches.



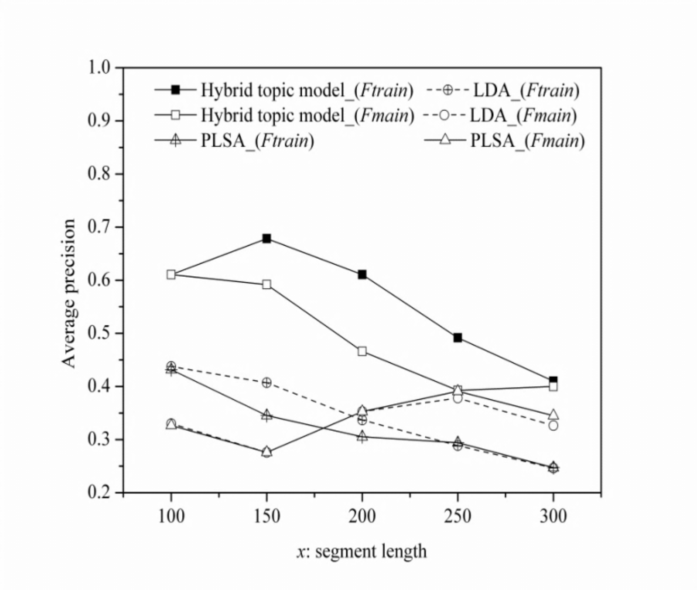


Figure 6. The average recall and precision values with different lengths of segments selected for motivation content extraction

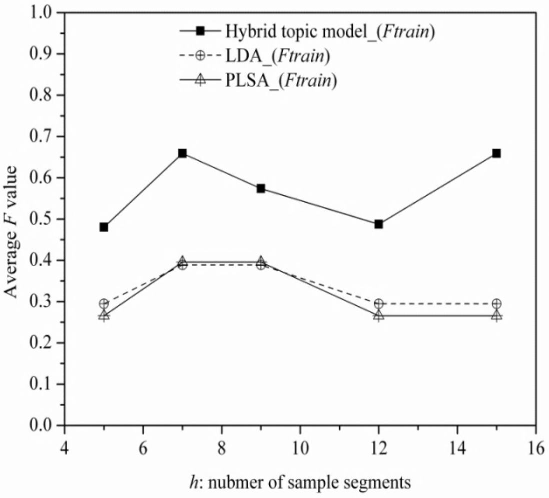
4.3 Influences of Parameters

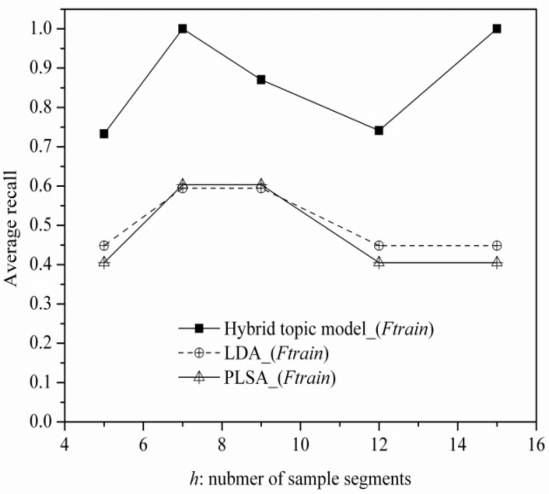
We also analyze the performance of our approach under different parameter settings. They include number of sample data, proportion factor β for sentence ranking, and number of topics for topic modeling and context modeling. These three parameters are important in our hybrid topic modeling, model learning and sentence ranking. In this experiment, we also rely on the task of extracting motivational aspect contents to study our approach.

4.3.1 Different number of sample data

This experiment evaluates the performance among the approaches using different quantities of sample data h. An approach which is able to achieve higher performance using fewer sample data is better. Since the approach can learn language patterns of the specified topical aspect with fewer samples, and as a result, the user can save time in preparing sample data. In the experiment, each sample segment $b_e \in B$ is selected from an individual document based on the human annotation data. It usually refers to one paragraph or several sentences. We test hybrid topic model_(F_{train}), LDA_(F_{train}) and PLSA_ (F_{train}) respectively.

Figure 7 summarizes the average performance of approaches for TSCE issue over difference number of sample data. Firstly, we can see that although different numbers of sample data are presented, our hybrid topic model can produce better results compared with other relevant approaches. It shows that our approach has the ability to learn language patterns from sample data for a specific topical aspect. In addition, we notice that using 7 to 10 samples enable these three approaches (i.e., our method, LDA and PLSA) to get relatively higher performance among using other numbers of sample data. It may suggest that using fewer samples (e.g., 5 and below) may not sufficiently represent the topical aspect of interest, while a larger number of samples (e.g., 15 samples and more) could also unnecessarily inject noise for both topic and context modeling.


Hybrid topic model_(*Ftrain*)
LDA_(*Ftrain*)
PLSA_(*Ftrain*)
Average *F* value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
4
6
8
10
12
14
16
h: nubmer of sample segments


Average recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Hybrid topic model_(*Ftrain*)
LDA_(*Ftrain*)
PLSA_(*Ftrain*)
4
6
8
10
12
14
16
h: nubmer of sample segments

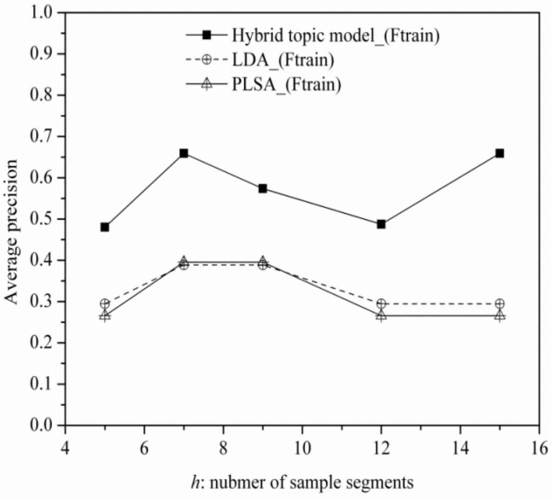

Hybrid topic model_(Ftrain)
LDA_(Ftrain)
PLSA_(Ftrain)
Average precision
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
4
6
8
10
12
14
16
h: nubmer of sample segments

Figure 7. The performance (average *F*, average precision and average recall) vs. different number of sample data for motivation content extraction

4.3.2 Proportion factor *β*

This experiment analyzes the performance of our approach with different settings of proportion factor β. Here we focus on investigate β in Equation (18) F_{train}, since our approach using F_{train} can obtain better results than using F_{main}. Factor β refers to the proportion for topic model and context model in sentence scoring and ranking, where topic number $K = T = 8$ and the sample number $E = 7$.

Table 3 shows the performance for motivation content extraction from patents using different settings of β. We start with $\beta = 0$. It means that we only use $P(w_j|c_l)$ in context models to score sentences. We notice that although only using the context model may not be able to generate some decent results, using context models in our approach can lead to comparable results compared to those using PLSA and LDA as shown in Section 4.2.1. To some extent, this shows that modeling a text collection from context space can be suggested as an alternative means to extract the contents of a specific topical aspect. When we increase the value of β to around 0.5, it shows that using the combination of topic model and context model can help to lift up the performance. When we switch to $\beta = 1$, meaning we only leverage $P(w_j|z_k)$ in topic model for sentence ranking, we notice that it still delivers high performance. It reveals that the proposed topic modeling, which suggests discriminative structures of topics, can achieve higher performance for TSCE than any others using PLSA and LDA shown in Section 4.2.1.

Table 3. The performance for motivation content extraction using our hybrid topic model_(F_{train}) based on different β

β	Average recall	Average precision	Average F value
0	0.7759	0.3782	0.5085
0.2	1.0000	0.4915	0.6591
0.4	1.0000	0.4915	0.6591
0.6	1.0000	0.4915	0.6591
0.8	1.0000	0.4915	0.6591
1	1.0000	0.4915	0.6591

4.3.3 Different number of topics

We also investigated the influence of topic number in our hybrid topic modeling. We also test our approach based on using F_{train} as the sentence ranking method. In order to test the performance under different number of topics, we set other parameters based on our findings in Section 4.3.1 and Section 4.3.2, i.e., the number of sample data $E = 7$ and the proportion factor $\beta = 0.4$.

Table 4 shows the performance of our approach using different number of topics in both topic modeling and context modeling for motivation content extraction. When we set $K = T <$ 5 topics, it shows that our hybrid topic model generates lower ROUGE-1 values. It may suggest that fewer topics in the hybrid topic modeling would have less power in modeling different topics and contexts of the text collection for topic-sensitive content extraction. When we increase the topic number from 6 to 10, we notice that the performance rises by about 10% compared to the results using less than 5 topics. It indicates that by switching to a higher topic number in both topic modeling and context modeling, it helps to better seize the characteristics of term distribution under different topics.

Table 4. The performance for motivation content extraction using our hybrid topic model_(F_{train}) based on different number of topics

K, *T*	Average recall	Average precision	Average *F* value
2	0.7328	0.3571	0.4802
4	0.7328	0.3571	0.4802
6	0.8707	0.4280	0.5739
8	1.0000	0.4915	0.6591
10	1.0000	0.4895	0.6572
12	0.8707	0.4280	0.5739
14	0.7414	0.3629	0.4873
16	0.7328	0.3571	0.4802

4.4 Scenario II: Extract Design Solutions and Reasons from Patents

The second scenario is to extract contents related to solution-reason aspect from a set of patent documents also intended for inject printer design. In this scenario, we can choose the parameter settings based on the results obtained in scenario I.

4.4.1 Performance among relevant approaches

For testing purpose, we also implement the relevant methods mentioned in Section 4.2 and conducted the similar experiments as in Section 4.2 based on different parameter combinations. Based on the observation in Section 4.2 and 4.3, we adopt hybrid topic model_(F_{train}) as our approach for solution and reason content extraction. In the evaluation, we did not count the segments shown in the claim section of patents when adopting ROUGE-1 measurement, since the claim section of patents were not included in the human annotation data. For a single document, each method returns the top x = 800 words to form the machine generated contents.

For comparison purpose, Table 5 shows the best results based on the average *F* value of each approach. Firstly, we notice that our hybrid topic model_(F_{train}) takes the lead again in all six approaches for the extraction of contents related to the solution and reason aspect. It gains about 0.6694 in the average *F* value. This result is about more than 2% higher than LDA_(F_{train})'s 0.6471, LDA_(F_{main})'s 0.6471 and PLSA_(F_{main})'s 0.6486, and about 4% higher than PLSA_(F_{train})'s 0.6232 and about 14% better than the baseline BL-s's 0.5293.

Table 5. The ROUGE-1 results for topic-sensitive content extraction associated with solution and reason aspect

	Average recall	Average precision	Average *F* value
Hybrid topic model_(F_{train})	**0.5865**	**0.7796**	**0.6694**
LDA_(F_{train})	0.5672	0.7530	0.6471
PLSA_(F_{train})	0.5460	0.7258	0.6232
LDA_(F_{main})	0.5672	0.7530	0.6471
PLSA_(F_{main})	0.5683	0.7554	0.6486
BL-s	0.4611	0.6213	0.5293

In addition, it is observed that all the approaches obtain better performance for the solution and reason content extraction than their performance for the motivation content extraction in Section 4.2. We have noticed that patent documents contain more contents about solution and reason aspect than contents associated with motivation aspect. If the resultant contents contain more words, then it is more likely to achieve higher values in ROUGE-1 measurement. Overall, experiments on these two scenarios demonstrate that our approach can generate acceptable results for topic-sensitive content extraction. Table 6 gives some example results of solution and reason aspect contents extracted by our approach.

Table 6. Topic combination for solution and reason aspect content extraction and some result examples extracted by our hybrid topic model_(F_{train})

Topic combination for solution and reason aspect content extraction

(The 25 most representative terms generated by our hybrid topic model_(F_{train}) for topic combination from topic model and the context model respectively. The terms are selected according to the probability $P(w|z)$ and $P(w|c)$.

z_{k_t} **in topic model**	c_{l_t} **in context model**
ink, cartridg, print, valv, suppli, pressur, form, head, chamber, printhead, nozzl, contain, liquid, surfac, includ, portion, reservoir, provid, posit, printer, embodi, claim, invent, seal, system	suppli, jet, droplet, ink, contain, eject, compris, receiv, chamber, respect, amount, flow, provid, plural, hold, us, color, black, drop, therein, fill, connect, includ, type, cartridg

Solution and reason aspect extracted from patent documents

(Some representative segments from the extracted results.)

ID 5	Color separation is provided in the printhead of the multicolor pen by staggering the individual color groups of nozzles in the scan direction while maintaining the same dot per inch spacing of the nozzles within the groups and between the groups as in the nozzle spacing in the single color printhead of the single color pen Construction of the multi-color ink jet pen is essentially the same as that of the single color pen especially as to size and external configuration and as to the pen mounting details in the pen carriage. Only changes are made in the interior of the ink reservoir to provide for the isolated storing of four colors of ink and to the nozzle substrate to provide the separate color channels for printing.
ID 13	In addition, because the joint portion 242 connected to the ink-droplet ejecting surface of the ink jet recording heads 220 is configured to be provided to the cover head 240, it is possible to perform the junction between the ink-droplet ejecting surface and the cover head 240 in such a manner that the respective nozzle lines 21A of the plurality of ink recording heads 220 are aligned with high precision to the cover head 240.
ID 19	The resulting effects include that a high quality monochrome image can be achieved by using a plurality of black ink compositions with different pigment concentrations, tone will be good in areas with low brightness, the gray balance will be stable, and variation will be reduced. Furthermore, the inventors have discovered that an image with favorable color balance can be achieved by using this yellow ink composition together with a specific color ink composition. Based on these findings, one advantage of the present invention is the ability to provide a yellow ink composition for inkjet recording that contains at least one pigment selected from a group consisting of C. I. Pigment Yellow 213, 185, and 155 as a colorant.

4.4.2 Performance under different lengths of contents

In this experiment, we also analyze the performance of our approach for solution and reason aspect content extraction when selecting contents with different lengths.

Figure 8 shows the average *F* values of approaches based on different lengths of segments selected as results for solution and reason contents. It shows that as the length of segments selected increases, our approach can help to extract contents that are more relevant to the topical aspect of solution and reason than other approaches. In addition, we notice that the average *F* values, shown in Figure 8, and the average recall values, shown in Figure 9, of those approaches increase linearly as the length of segment increases. In our case, the contents of solution and reason aspect annotated usually contain more than four times of words as those of motivation aspect, therefore, a higher chance of matching with human annotation if a longer segment is presented.

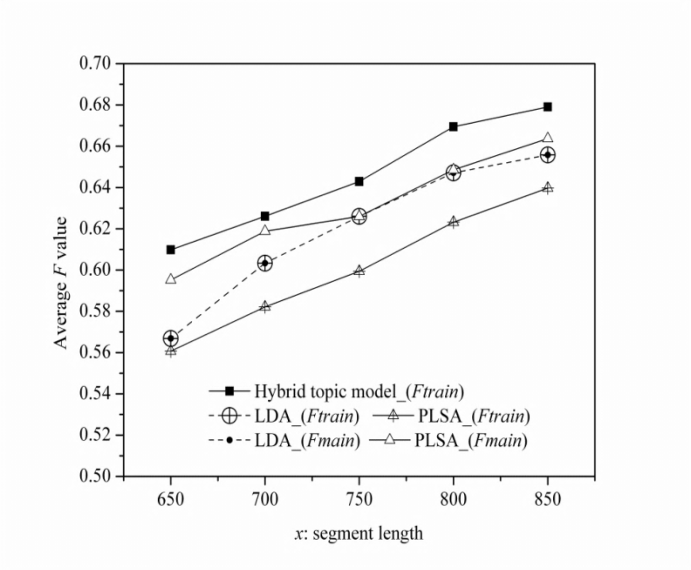
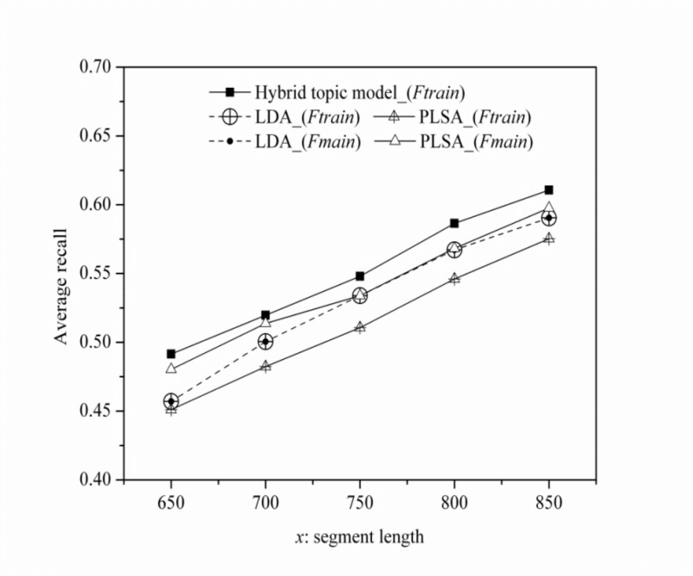


Figure 8. The average *F* values with different lengths of segments selected for solution and reason content extraction

Figure 9 shows the average recall and precision values with respect to different lengths of segments selected as solution and reason contents. We observe that with the same length of segments, our approach can achieve both higher recall and precision performance compared to other four relevant approaches. The higher average recall reveals that our approach can possibly generate more relevant contents. In addition, as the length increases, it shows that our approach can also deliver higher average precision. The experiments of these two cases demonstrate that by integrating the intrinsic differences of topic and context in modeling, the proposed hybrid topic model can better uncover hidden topics and their aspects for TSCE problems.



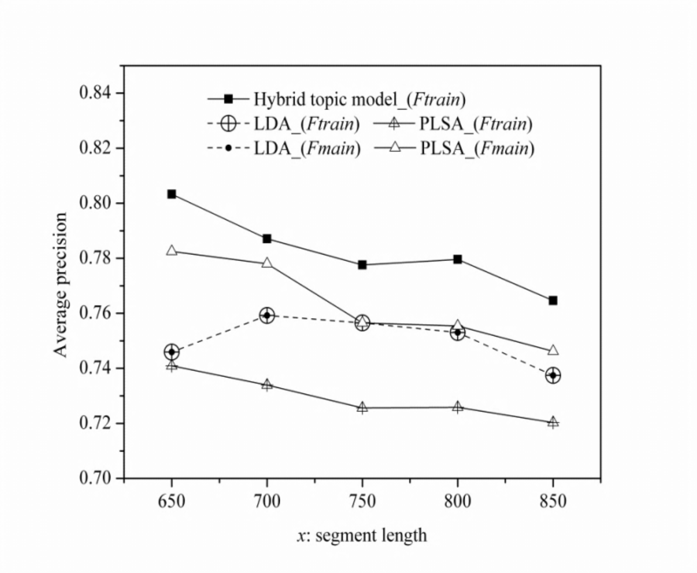


Figure 9. The average recall and precision values with different lengths of segments selected for solution and reason content extraction

5. Conclusion

This paper reports our study on an information extraction problem named topic-sensitive content extraction. Given a document collection, TSCE takes sample segments that are specified as the instances of a topical aspect as inputs. It aims at discovering contents that address the specified topical aspect from each individual document in the collection. This is helpful when users intend to seek for information pertaining to a specific subtopic, e.g., design issues, solutions and cause-effect. In order to tackle TSCE, we have proposed a new hybrid topic model that takes into account topic modeling and context modeling in learning language features for the specified topical aspect. Specifically, the proposed model leverages term profiles to model topic contexts and explores local and global differences between topics in the topic network. As a result, the proposed hybrid topic modeling approach can offer more discriminative power to recognize contents related to the specified aspect. The experiments of extracting contents closely associated with the motivation aspect and the solution-reason aspect from patent documents demonstrate the merits of the proposed approach when compared to several prevailing baseline approaches.

Our key contribution of the present paper is a systematic approach for the question of extracting topic-sensitive contents from textual document, which received less attention in relevant studies. Our finding highlights the importance of using context and local features in suggesting different topical aspects in a single document. In addition, the context modeling by incorporating the prior knowledge into topic modeling to some extent is able to suggest discriminative structures of topics. Moreover, by learning the language features from both topic space and context space, it is able to achieve higher performance to extract contexts relevant to the topical aspect indicated by a user.

There are still rooms to improve the studies of topic-sensitive content extraction. Firstly, our approach is for documents relevant to a domain. It is unclear as to how the performance would be different for different granularity of relevant document set. In addition, we only use the context information in the neighborhood region. It could be better to explore other context, such as design graphs and document structures, in pattern learning and topical aspect modeling.

Acknowledgments

The authors would like to express their thanks to Dr. Peter A from UK IPO for the insights on patent informatics and helping examining experimental results.

References

Blei, David M. , Ng, Andrew Y. , & Jordan, Michael I. . (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993-1022.

Cai, Deng, Mei, Qiaozhu, Han, Jiawei, & Zhai, Chengxiang. (2008). *Modeling hidden topics on document manifold*. Paper presented at the Proceedings of the 17th ACM conference on Information and knowledge management, Napa Valley, California, USA.

Cai, Deng, Wang, Xuanhui, & He, Xiaofei. (2009). *Probabilistic dyadic data analysis with local and global consistency*. Paper presented at the Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, Quebec, Canada.

Dempster, Arthur, Laird, Nan, & Rdin, Donald. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal statistical society, Series B, 39*(1), 1-38.

Hofmann, Thomas (1999). *Probabilistic latent semantic indexing*. Paper presented at the Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, Berkeley, California, United States.

Jiang, Jing , & Zhai, Chengxiang (2006). Extraction of coherent relevant passages using hidden Markov models. *ACM Transactions on Information Systems, 24*(3), 295-319.

Jindal, Nitin, & Liu, Bing. (2006). *Mining Comparative Sentences and Relations*. Paper presented at the Proceedings of The 21 National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, Boston, Massachusetts, USA.

Karen, Spack Jones. (2007). Automatic summarising: The state of the art. *Information Processing & Management, 43*(6), 1449-1481.

Li, Jiwei, Li, Sujian. (2013). A Novel Feature-based Bayesian Model for Query Focused Multi-document Summarization. *Transactions of the Association for Computational Linguistics*, 1, 89-98.

Liang, Yan, Liu, Ying, Lee, Wing Bun, & Kwong, Chun Kit. (2011). *Building a Semantic Graph based on Sequential Language Model for Topic-Sensitive Content Extraction*. Paper presented at the Proceedings of the 9th Workshop on Mining and Learning with Graphs in conjunction with SIGKDD 2011, San Diego, CA, USA.

Lin, Chin-Yew, & Hovy, Eduard. (2003). *Automatic evaluation of summaries using N-gram co-occurrence statistics*. Paper presented at the Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Edmonton, Canada.

Liu, Zhiyuan, Huang, Wenyi, Zheng, Yabin, & Sun, Maosong. (2010). *Automatic keyphrase extraction via topic decomposition*. Paper presented at the Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, Massachusetts.

Lu, Hsin-Min. (2015). Detecting short-term cyclical topic dynamics in the user-generated content and news. *Decision Support Systems, 70*, 1-14.

Lu, Yue, Castellanos, Malu, Dayal, Umeshwar, & Zhai, ChengXiang. (2011). *Automatic construction of a context-aware sentiment lexicon: an optimization approach*. Paper presented at the Proceedings of the 20th international conference on World wide web, Hyderabad, India.

Mavridis, Themistoklis, Symeonidis, Andreas L. (2014). Semantic analysis of web documents for the generation of optimal content. *Engineering Applications of Artificial Intelligence*, 35, 114-130.

McLachlan, Geoffrey J., & Krishnan, Thriyambakam. (2008). *The EM algorithm and extensions (Second Edition)*. USA: Wiley-Interscience.

Niu, Lingfeng, & Shi, Yong. (2010). *Semi-supervised PLSA for Document Clustering*. Paper presented at the Proceedings of the 2010 IEEE International Conference on Data Mining Workshops, Sydney, NSW.

Pang, Bo, & Lee, Lillian. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135.

Radev, Dragomir R., Jing, Hongyan, Styś, Małgorzata, & Tam, Daniel. (2004). Centroid-based summarization of multiple documents. *Information Processing &amp; Management, 40*(6), 919-938.

Singhal, Amit. (2001). Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 24*(4), 35-42.

Tang, Huifeng, Tan, Songbo, & Cheng, Xueqi. (2009). A survey on sentiment detection of reviews. *Expert Systems with Applications, 36*(7), 10760-10773.

Tang, Jian, Zhang, Ming, & Mei, Qiaozhu. (2013). *One theme in all views: Modeling consensus topics in multiple contexts*. Paper presented at Proceedings of the 2013 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA.

Wang, Dingding, Zhu, Shenghuo, Li, Tao, & Gong, Yihong. (2009). *Multi-document summarization using sentence-based topic models*. Paper presented at the Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Suntec, Singapore.

Wang, Hongning, Lu, Yue, & Zhai, Chengxiang. (2010). *Latent aspect rating analysis on review text data: a rating regression approach*. Paper presented at the Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, DC, USA.

Wang, Mengqiu, & Si, Luo. (2008). *Discriminative probabilistic models for passage based retrieval*. Paper presented at the Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, Singapore, Singapore.

Xue, Gui-Rong, Dai, Wenyuan, Yang, Qiang, & Yu, Yong. (2008). *Topic-bridged PLSA for cross-domain text classification*. Paper presented at the Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, Singapore, Singapore.

Yao, Liang, Zhang, Yin, Chen, Qinfei, Qian, Hongze, Wei, Baogang, Hu, Zhifeng. (2017). Mining coherent topics in documents using word embeddings and large-scale text data. *Engineering Applications of Artificial Intelligence*, 64, 432–439.

Zhan, Jiaming, Loh, Han Tong, & Liu, Ying. (2009). Gather customer concerns from online product reviews - A text summarization approach. *Expert Systems with Applications, 36*(2, Part 1), 2107-2115.

Zhao, Wayne Xin, Jiang, Jing, He, Jing, Song, Yang, Achananuparp, Palakorn, Lim, Ee-Peng, & Li, Xiaoming. (2011). *Topical keyphrase extraction from Twitter*. Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, Portland, Oregon.