Engineering Applications of Artificial Intelligence 70 (2018) 81–91

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Extracting topic-sensitive content from textual documents—A hybrid topic model approach

Yan Liang [a], Ying Liu [a,*], Chong Chen [a], Zhigang Jiang [b]

[a] Institute of Mechanical and Manufacturing Engineering, School of Engineering, Cardiff University, Cardiff CF24 3AA, UK
[b] College of Machinery and Automation, Wuhan University of Science & Technology, Wuhan 430081, China

ARTICLE INFO

Keywords:
Topic-sensitive content
Probabilistic topic modeling
Topic network
Semi-supervised

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.

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 et al., 2011) and discovery of user ratings for product aspects (Wang et al., 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

* Corresponding author.
E-mail address: LiuY81@cardiff.ac.uk (Y. Liu).

https://doi.org/10.1016/j.engappai.2017.12.010
Received 10 September 2016; Received in revised form 15 October 2017; Accepted 18 December 2017
0952-1976/

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

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 et al., 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 and 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 and Zhai, 2006; Wang and 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 et al., 2004). Some other research efforts provide users with sentiment-level information, such as review sentiment detection (Jindal and Liu, 2006; Tang et al., 2009) and opinion summarization (Pang and Lee, 2008; Zhan et al., 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 and Li (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 et al. (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 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