Expert Systems With Applications 60 (2016) 27–38

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Concept over time: the combination of probabilistic topic model with wikipedia knowledge

Liang Yao, Yin Zhang*, Baogang Wei, Lei Li, Fei Wu, Peng Zhang, Yali Bian

College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Article history:
Received 15 December 2014
Revised 4 March 2016
Accepted 14 April 2016
Available online 29 April 2016

Keywords:
Concept over time
Topic model
Wikipedia knowledge

ABSTRACT

Probabilistic topic models could be used to extract low-dimension aspects from document collections, and capture how the aspects change over time. However, such models without any human knowledge often produce aspects that are not interpretable. In recent years, a number of knowledge-based topic models and dynamic topic models have been proposed, but they could not process concept knowledge and temporal information in Wikipedia. In this paper, we fill this gap by proposing a new probabilistic modeling framework which combines both data-driven topic model and Wikipedia knowledge. With the supervision of Wikipedia knowledge, we could grasp more coherent aspects, namely, concepts, and detect the trends of concepts more accurately, the detected concept trends can reflect bursty content in text and people's concern. Our method could detect events and discover events specific entities in text. Experiments on New York Times and TechCrunch datasets show that our framework outperforms two baselines.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

The explosion of online text content, such as news, blogs, Twitter messages and QA community messages has given rise to the challenge to understand the very dynamic sea of text. To address the challenge, we need to extract the concepts from the sea of text, for the reason that "Concepts are the glue that holds our mental world together" (Murphy, 2002) and "Without concepts, there would be no mental world in the first place" (Bloom, 2003). Besides the content of concepts, we also need to mine the dynamic patterns of concepts, in order to reflect the dynamic nature of the large text data sets.

A lot of text mining tasks, especially aspects extraction tasks, employ statistical topic models such as PLSA (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). However, these unsupervised models without any human knowledge often result in topics that are not interpretable, i.e., could not generate semantically coherent *concepts* (Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009; Mimno, Wallach, Talley, Leenders, & McCallum, 2011).

Many of the large data sets do not have *static* patterns, they are *dynamic* instead. The data are often collected over time and have time stamps. Concepts rise and fall as time goes. For instance, the concept "Earthquake" bursts in China's news and social networks in May 2008, in August, "Earthquake" falls, by contrast, the concept "Olympic Games" rises.

In recent years, some knowledge-based topic models (Andrzejewski & Zhu, 2009; Andrzejewski, Zhu, & Craven, 2009; Andrzejewski, Zhu, Craven, & Recht, 2011; Chemudugunta, Holloway, Smyth, & Steyvers, 2008; Chen et al., 2013a; 2013b; Doshi-Velez, Wallace, & Adams, 2015; Yao, Zhang, Wei, Qian, & Wang, 2015) and dynamic topic models (Blei & Lafferty, 2006; Kalyanam, Mantrach, Saez-Trumper, Vahabi, & Lanckriet, 2015; Wang & McCallum, 2006; Wei, Sun, & Wang, 2007; Yan, Guo, Lan, Xu, & Cheng, 2015) have been proposed to overcome above issues. However, they could not process concept knowledge and temporal information in Wikipedia which are important side information which could be used for topic modeling.

To overcome this shortcoming, we propose a new probabilistic framework, called *Concept over Time*, which combines topic modeling techniques and Wikipedia knowledge, in particular LDA-style topic model and Wikipedia entries with their view logs. The proposed framework explicitly models text content with time, Wikipedia knowledge and Web users' behaviors, could extract more accurate dynamic patterns with more specific and coherent entities.

* Corresponding author.
E-mail addresses: yaoliang@zju.edu.cn (L. Yao), yinzh@zju.edu.cn (Y. Zhang), wbg@zju.edu.cn (B. Wei), roylee@zju.edu.cn (L. Li), wufei@zju.edu.cn (F. Wu), pengzhang1991@zju.edu.cn (P. Zhang), bianyali@zju.edu.cn (Y. Bian).

http://dx.doi.org/10.1016/j.eswa.2016.04.014
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

The contributions of the paper are threefold: (1) It proposes a novel knowledge-based dynamic topic model based on Wikipedia knowledge and Web users' behaviors. (2) It provides Gibbs sampling inference methods which could handle Wikipedia knowledge and Web users' behaviors properly. (3) Comprehensive experimental results on two datasets demonstrate our method outperforms two state-of-the-art topic models.

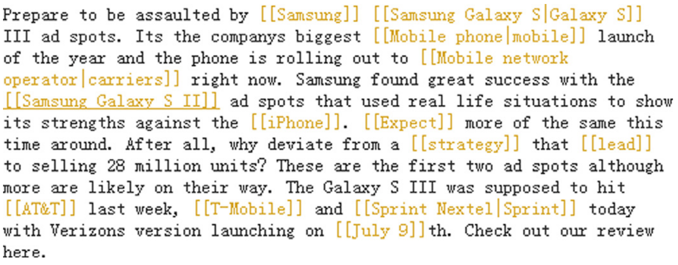
We begin this paper by reviewing some related works, including studies which devote to improving the interpretability of topic models, mainly by incorporating domain knowledge into topic models, studies which focus on the dynamic topic models, studies which represent text in a more semantically explicit way than bag of words, and studies which utilize Wikipedia knowledge in nature language processing. In the remainder of the paper, we first describe our framework, then conduct experiments on real data sets, and analyze experimental results, which show the effectiveness of our method. Finally, some discussions and conclusions are given.

2. Related work

To overcome the shortcoming of topic coherence in topic models, especially in LDA, some previous studies incorporate domain knowledge into LDA model. Andrzejewski and Zhu (2009) proposed topic-in-set knowledge which restricts topic assignment of terms to a subset of topics. They improved the topic-in-set knowledge in (Andrzejewski et al. (2011)) by incorporating general knowledge specified by first-order logic. Similarly, Concept model and Concept-topic model were proposed by using ontologies like Open Directory Project (ODP) or The Cambridge International Dictionary of English (CIDE) (Chemudugunta et al., 2008). The DF-LDA model in (Andrzejewski et al. (2009)) can handle knowledge in the form of must-links and cannot links. A must-link states that two words should belong to the same topic, while a cannot-link states that two words should not be in the same topic. Chen et al. (2013b) presented LDA with Multi-Domain Knowledge (MDK-LDA) which can use prior knowledge from multiple domains. In (Chen et al. (2013a)), MC-LDA (LDA with must-link set and cannot-link set), was proposed as an extension of MDK-LDA. Lately, (Doshi-Velez et al., 2015) proposed a strategy for achieving interpretability by exploiting controlled structured vocabularies in which words are organized into tree-structured hierarchies. Probase-LDA (Yao et al., 2015), a method that combines topic model and a probabilistic knowledge base was put forward recently. Despite these methods utilize knowledge in many ways, they do not consider concept knowledge in Wikipedia, which is an important knowledge form for both human beings and machines.

To measure the coherence of topic models, (Mimno et al., 2011) presented an automatic coherence measure of topic models, which automates the human judge method in (Chang et al. (2009)). At the same time, they put forward an unsupervised method which improves the coherence score by considering the word co-occurrence in a corpus. Chuang, Gupta, Manning, and Heer (2013) measured the correspondence between a set of latent topics and a set of reference concepts when applying topic model to domain-specific tasks, which provides another way to measure the coherence.

Some dynamic topic models have been proposed to mine dynamic patterns, e.g., Dynamic Topic Models (DTMs) (Blei & Lafferty, 2006), Topic over Time (TOT) (Wang & McCallum, 2006) and Dynamic Mixture Model (DMM) (Wei et al., 2007). These models all take time into consideration. In DTMs, topic evolutions are modeled through collections sliced into certain periods of time. In TOT, continuous time stamps are put into LDA-style topic model as observations. In DMM, the evolution is captured by modeling the dependency between two consecutive documents. Although these dynamic topic models use time information of documents, they may not clearly reflect what people care about in a certain period. Recently, Kalyanam et al. (2015) use social context information in Twitter to detect topic evolution in news articles. Yan et al. (2015) exploit the burstiness of biterms in microblog as prior knowledge and incorporate it into a short text topic model. Our method, on the other hand, can reflect people's concern by utilizing Wikipedia article traffic statistics.



Prepare to be assaulted by [[Samsung]] [[Samsung Galaxy S|Galaxy S]]
III ad spots. Its the companys biggest [[Mobile phone|mobile]] launch
of the year and the phone is rolling out to [[Mobile network
operator|carriers]] right now. Samsung found great success with the
[[Samsung Galaxy S II]] ad spots that used real life situations to show
its strengths against the [[iPhone]]. [[Expect]] more of the same this
time around. After all, why deviate from a [[strategy]] that [[lead]]
to selling 28 million units? These are the first two ad spots although
more are likely on their way. The Galaxy S III was supposed to hit
[[AT&T]] last week, [[T-Mobile]] and [[Sprint Nextel|Sprint]] today
with Verizons version launching on [[July 9]]th. Check out our review
here.

Fig. 1. An example of a bag of Wikipedia articles in TechCrunch dataset.

Studies for Explicit Semantic Analysis (ESA) (Gabrilovich & Markovitch, 2007) focus on finding explicit rather than latent topics. Instead of using a bag of words to represent a topic, the method uses a bag of Wikipedia articles, or a distribution over the entire set of Wikipedia articles, to represent a topic. Besides Explicit Semantic Analysis, many natural language processing applications exploit Wikipedia knowledge. Chenthamarakshan, Melville, Sindhwani, and Lawrence (2011) use Wikipedia articles and categories to label texts with concepts. Ferschke, Daxenberger, and Gurevych (2013) introduced how revisions in Wikipedia could be used in spelling error correction, vandalism detection and so on. Tinati, Tiropanis, and Carr (2013) demonstrated that Wikipedia can provide a useful measure for detecting social trends and events using the Wikipedia article view logs. Georgescu et al. (2013) presented a web-based system that supports the entity-centric, temporal analytics of event-related information in Wikipedia by analyzing the whole history of article updates. Inspired by these, we use Wikipedia articles and view logs for better topic modeling.

3. The proposed framework

3.1. Represent document as a bag of wikipedia articles with view logs

We first map text to Wikipedia articles. We exploit the Wikify service in Wikipedia miner (Milne & Witten, 2013). The approach first gathers all n-grams for a document and retains those whose link probability exceeds a certain threshold. The link probability of a phrase is defined as the number of Wikipedia articles that use it as an anchor, divided by the number of articles that mention it at all. Then the approach balances the commonness of a sense with its relatedness to the surrounding context to disambiguate detected phrases to ensure that they link to the appropriate article.

After Wikification, a list of Wikipedia articles with their link probability p_{di} (the balance of commonness and relatedness) to a disambiguated article are returned. The larger the link probability is, the more likely that the term is important in the input text. As an example, a document in TechCrunch dataset has been annotated in Fig. 1, and the link probabilities of detected articles are shown in Fig. 2, the size of articles means the link probability.

Next, we exploit the edit time stamp of text to attain the view logs of Wikipedia articles returned above. If an article is linked to text, we get the number of article views from Wikipedia article traffic statistics.[1]

[1] http://stats.grok.se/.

Download English Version:

https://daneshyari.com/en/article/383104

Download Persian Version:

https://daneshyari.com/article/383104

Daneshyari.com