Neurocomputing 242 (2017) 40–50

Contents lists available at ScienceDirect

Neurocomputing

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

Semantic-based topic detection using Markov decision processes

Qian Chen[a,b], Xin Guo[a,*], Hexiang Bai[a]

[a] School of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi 030006, PR China
[b] School of Electronics and Information Engineering, Tongji University, Shanghai 201804, PR China

ARTICLE INFO

Article history:
Received 24 July 2015
Revised 9 January 2017
Accepted 7 February 2017
Available online 21 February 2017

Communicated by Huaping Liu

Keywords:
Community discovery
Markov Decision Process
Topic detection
Topic graph
Topic pruning

ABSTRACT

In the field of text mining, topic modeling and detection are fundamental problems in public opinion monitoring, information retrieval, social media analysis, and other activities. Document clustering has been used for topic detection at the document level. Probabilistic topic models treat topics as a distribution over the term space, but this approach overlooks the semantic information hidden in the topic. Thus, representing topics without loss of semantic information as well as detecting the optimal topic is a challenging task. In this study, we built topics using a network called a topic graph, where the topics were represented as concept nodes and their semantic relationships using WordNet. Next, we extracted each topic from the topic graph to obtain a corpus by community discovery. In order to find the optimal topic to describe the related corpus, we defined a topic pruning process, which was used for topic detection. We then performed topic pruning using Markov decision processes, which transformed topic detection into a dynamic programming problem. Experimental results produced using a newsgroup corpus and a science literature corpus showed that our method obtained almost the same precision and recall as baseline models such as latent Dirichlet allocation and KeyGraph. In addition, our method performed better than the probabilistic topic model in terms of its explanatory power and the runtime was lower compared with all three baseline methods, while it can also be optimized to adapt the corpus better by using topic pruning.

1. Introduction

Due to the rapid development of computer networks and social media, the volumes of various types of data have been increasing rapidly, especially user-generated content. Therefore, there is an urgent need to discover interesting patterns hidden in these massive volumes of data. In this study, we focused on text data because texts are generated in natural human language and the semantic information hidden per unit size in text is richer than that in other data formats such as video, images, and audio. We aimed to discover latent hierarchical structures called topics in large-scale corpora by topic detection.

Topic detection was initiated in the topic detection and tracking (TDT) research program early in 1998, which aimed to discover topics or trends in various type of online media text data. TDT has attracted much attention in the last two decades in many application areas, such as online reputation monitoring [6], public opinion detection [7], and user interest modeling [18]. Topic detection is a fundamental application area in the text mining community, including text classification and clustering, information retrieval, and document summarization. [1]. Topic detection plays an important role in information retrieval and data mining, and it is an effective tool for organizing and managing text data such as newswire archives and research literature.

Unlike other existing applications in text mining and information retrieval, topic detection is an entirely unsupervised learning task without any topic classes or structure labels. In general, a topic is represented as related sets of keywords, and thus important descriptions can be given to topics or events. Many text clustering algorithms that typically compute similarities have been developed for topic detection, such as single pass incremental clustering algorithms [2] and incremental clustering algorithms [10]. Since the latent Dirichlet allocation (LDA) method was proposed by Blei in 2003 [4], the probabilistic topic model (pTM) has attracted much attention in the fields of information retrieval, text mining, and other areas. Essentially, pTM is a type of probabilistic model used for topic modeling, including LSA, pLSA, LDA, and various extension versions of pTM, which treat a topic as a distribution over the term space.

Despite the success of pTM, it has several drawbacks, as follows. (1) The inference algorithm used in the model can be too complex and much time is required to generate the topic word

* Corresponding author.
E-mail addresses: chenqian@sxu.edu.cn (Q. Chen), guoxinjsj@sxu.edu.cn (X. Guo), baihx@sxu.edu.cn (H. Bai).

http://dx.doi.org/10.1016/j.neucom.2017.02.020
0925-2312/

distribution, especially for large noisy and unbalanced corpora such as social media data. (2) There is a lack of explanatory power because the methods mentioned above ignore the semantic relationships between terms as well as topics. Existing explicit semantic topic detection methods usually build an ontology or some other structure containing rich semantic information, before employing ontology mapping, calculating, and reasoning to compute the similarity among terms to identify semantic relationships and facilitate semantic-based topic detection. However, building a general ontology requires a long time, despite the relatively low workload of building domain ontology. (3) Most importantly, to the best of our knowledge, topic optimization is not considered in the topic detection algorithm, which aims to optimize the topics generated and select appropriate topic words. Therefore, designing a new topic detection method that considers semantics and automatically selects the optimal topic set with low complexity in terms of time and space is a new challenge. In this study, we investigated the importance of topic semantic explicability and topic optimization, and we developed a topic graph establishment method, which represent topics using a network, where topics are represented as concept nodes and their semantic relationships using WordNet. In order to find the optimal topic that describes the related corpus, we define a topic pruning process and perform topic pruning using Markov decision processes (MDPs).

After completing our study, we recently found that Sayyadi and Raschid [11] proposed a graph analytical approach for topic detection by representing a topic as a graph based on keyword co-occurrence, as in our proposed method. However, there are two differences: (1) in our topic representation, we focus mainly on semantic information using an external knowledge-base; and (2)we propose a topic pruning process based on Markov decision processes, whereas Sayyadi and Raschid [11] did not consider topic optimization. Nevertheless, the conclusion of Sayyadi and Raschid [11] that word co-occurrence can obtain superior runtime performance compared with other solutions demonstrates that a similar approach can outperform pTM in terms of its lower time complexity.

In our proposed method, we first abstract the topics using a novel network called a topic graph, where the topics are represented as concept nodes and their semantic relationships using the WordNet database. Second, in order to find the optimal topic that describes the related corpus, we define a topic pruning processes, which is then used for topic detection. Third, we perform topic pruning using MDPs, which transforms topic detection into a dynamic programming problem. The main contributions of this study are summarized as follows.

(1) We propose a novel graphical representation for topics, which can identify related concept nodes as well as considering the relationships between concept nodes to detect deep semantic information hidden in the topics.
(2) We define a drill-down operator and we perform topic pruning using MDPs, thereby transforming topic detection into a dynamic programming problem, and thus the optimized topics can be adapted better to the corpus.
(3) We annotated the NIPS12 corpus, which include 1740 articles, and we also evaluated our approach using two different categories of corpus, i.e., newsgroup100 and NIPS12, in terms of the precision and recall, where the experiment results verified the efficiency of our approach.

The remainder of this paper is organized as follows. Related research is introduced in Section 2. We formulate the problem in a formal manner in Section 3 and Section 4 explains the topic graph construction process. We define topic optimization in Section 4 and the topic pruning algorithm is described in Section 5. Section 6 presents the details of our experiments and performance evaluations. Finally, we give our conclusions in Section 7.

2. Related work

In general, topic detection can be divided into two modes: online and off-line. Online topic detection aims to discover dynamic topics over time as new topics appear. Many studies have focused on new approaches to event detection, novel topic discovery, online topic evolution, and other problems in the online mode, which requires an incremental algorithm. Off-line topic detection is also known as retrospective topic/event detection, and it treats all documents in a corpus as a batch, before detecting topics one at a time [20]. In this study, we focused mainly on the off-line mode. Topic detection methods can be categorized according to three types: document clustering-based topic detection, pTM-based topic detection, and graph-based topic detection.

In document clustering-based topic detection, each document is represented as a vector using TF-IDF or improved TF-IDF, and each topic is simply a set of keywords. Brants proposed a variation of TF-IDF for detecting topics [19]. Many studies have considered retrospective topic detection using document clustering, including the well-known augmented group average clustering (GAC) method [20].

The LDA model is a Bayesian hierarchical probabilistic generative model, which was first proposed by Blei et al. [4]. In this method, each document is modeled as a discrete distribution over topics, and each topic is regarded as a discrete distribution over terms. LDA is used widely in text mining and other fields, and it is regarded as a powerful tool for topic modeling. The original LDA method used a variational expectation maximization (VEM) algorithm to infer topics for LDA [4], but stochastic sampling inference based on Gibbs sampling was proposed by Steyvers and Griffiths [12] for LDA. Similar to Sayyadi and Raschid [11], we denote LDA-GS as LDA with Gibbs sampling and LDA with VEM as LDA-VEM.

The two types of topic detection methods mentioned above only consider words, especially in the LDA model, where words are generated conditionally independent of a given distribution. In fact, there are richer relationships between words. Graph-based topic detection methods focus on between-words relationships. The co-occurrence patterns between words were considered in previous studies. For example, Petkos [3] treated the topic detection problem as a frequent pattern mining problem and proposed a soft frequent pattern mining algorithm. Cataldi also built a co-occurrence graph for tweets with an extra temporal dimension [21]. Sayyadi and Raschid proposed a graph analytical approach for topic detection called KeyGraph (KG) [11]. KG is essentially a keyword co-occurrence graph based on an off-the-shelf community detection algorithm for grouping co-occurring keywords into communities. Each community was a constellation of keywords representing a topic. Inspired by KG, we use betweenness-metric-based community detection for topic extraction in our proposed method.

Many studies have aimed to extend the LDA model. Some extended LDA models have been used to model authorship information [22], while others aim to capture the most recent language usage for sentiments and topics [23]. The biterm topic model is applied to short texts such as tweets based on an extension of the LDA [24]. The inverse regression topic model combines metadata with the LDA to utilize structural information in each document [25]. The correlated topic model is used to model the correlations between topics to remove the assumption of independence in the LDA [16]. Recently, deep learning techniques have been used to obtain low-dimensional representations of word and documents by word embedding. Thus, anovel neural topic model [15] was proposed to combine the advantages of topic models and neural networks, but it is essentially a supervised learning model. In pTM,