Optik 125 (2014) 2164–2169

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

Topic tracking with Bayesian belief network☆

Jian-min Xu[a,*], Shu-fang Wu[a,b], Yu Hong[c]

[a] Hebei University College of Management, Hebei Baoding 071002, China
[b] Hebei Software Institute Department of Information Engineering, Hebei Baoding 071000, China
[c] Soochow University School of Computer and Technology, Suzhou 215006, China

ARTICLE INFO

Article history:
Received 12 May 2013
Accepted 11 October 2013

Keywords:
Bayesian belief network
Topic tracking
Static topic model
Dynamic topic model

ABSTRACT

The task of topic tracking is to monitor a stream of stories and find all subsequent stories that discuss the same topic. Using Bayesian belief network we give three topic tracking models: a static topic model BSTM and two dynamic topic models BDTM-I, BDTM-II. BDTM-II merges the advantages of BSTM and BDTM-I, has better tracking performance than the former two, and effectively alleviates topic drift phenomenon. Applying unrelated coming stories to update BDTM-I and BDTM-II can filter noises existed in topics. Experiments on TDT corpora show that BSTM decreases $(C_{det})_{norm}$ by 5.5% comparing to VSM, BDTM-II decreases $(C_{det})_{norm}$ by 6.3% and 6.0% comparing to BSTM and BDTM-I respectively, using unrelated stories can improve the tracking performance.

1. Introduction

Topic detection and tracking (TDT) is a new line research composed of three major sub-problems [1–3]: splitting a continuous stream of news into stories (segmentation), gathering stories into groups that each discuss a single topic (detection), and exploiting user feedback to monitor a stream of news for additional stories on a specified topic (tracking). This last task, topic tracking, is the focus of our research in this paper.

This paper presents four contributions to the field of topic tracking: the first is to give a static topic model [4] based on Bayesian belief network model (BSTM), including both topology and probability computation. The second is to provide a dynamic topic model [4,5] BDTM-I, which updates the topic model by related stories. Comparing BSTM to BDTM-I, we find that BSTM lowers False Alarm Probability [6] but improves Miss Probability [6], BDTM-I lowers Miss Probability but improves False Alarm Probability. The third contribution is to solve this problem by giving another dynamic topic model BDTM-II. Finally, We give a quantitative method to update dynamic models by both related and unrelated stories.

To evaluate the tracking system, the 2003 evaluation method C_{det} [6,7] is used. Experiments show that BSTM and BDTM-I has better tracking performance than VSM (Vector Space Model), BDTM-II has better performance than BSTM and BDTM-II.

2. Related works

This section includes three parts: the first part introduces research status, the second part gives some definitions used in our paper, the third part introduces basic Bayesian belief network model.

2.1. Research status

Early research on topic tracking did not focus on topic form and modeling method, but dedicated to the transplantation of related technologies, such as applying information extracting, filtering and classification technology to topic tracking. Watanabe et al. [8] explained topic tracking as a process of information extraction. Zhang and Callan [9] used information filtering technology based on content to block unrelated stories. But there was difference between topic tracking and information processing, for example temporal factor was considered in topic tracking not in information processing, so information processing technology was not completely suitable for topic tracking. Using the existing classification methods, such as KNN [10], D-tree [11], and linear classifier [12] to classify news stream into two categories was another emphasis on topic tracking research in this time. Because of the existing topics was identified by a few samples (1–4), there was limitations when classification technologies was used in topic tracking.

The prospective research of topic model was from Allan et al. [13], who applied Vector Space Model (VSM) used in information retrieval to describe the feature space of topic, Yang and Allan [14–17] continued to improve vector space model by different strategies. At the same time, another modeling method based on language model [18,19] was appearing. Lavrenko and

☆ Supported by China Postdoctoral Science Foundation under Grant No. 20070420700; Science Foundation of HeBei Province under Grant No. F2011201146.
* Corresponding author at: Hebei University College of Management, Hebei Baoding 071002, China.
E-mail address: shufang_44@126.com (J.m. Xu).

0030-4026/$ – see front matter © 2013 Elsevier GmbH. All rights reserved.
http://dx.doi.org/10.1016/j.ijleo.2013.10.044

Croft [20] proposed a relevance model, Nallapai [21] established a topic description based on semantic language model, Hong et al. [22] gave a semantic domain based language model used chapter structure and dependencies, Ramage et al. [23] applied LDA [24] to topic model. Language model could be explained as probability model [25] as the computation process referring to conditional probability. Bayesian belief network model [26] as the extending of probability model has been successfully used in information retrieval during the past several decades [27]. Our research is to use Bayesian belief network model to achieve topic tracking.

2.2. Relevant definition

In order to distinguish the concepts in linguistic, DTD evaluation conference gives the definition of topic and other general concepts [15].

Definition 1. *Topic*: A topic is defined to be a seminal event or activity, along with all directly related events and activities.

Definition 2. *Story*: A story is closely related to a topic, including two or more news segmentations to independent describe an event.

In addition to the above definitions, we also use the following definitions in the paper [4]:

Definition 3. *Static topic mode (STM)*: STM emphasizes the conservation of initial seminal contents and will not update in the process of topic tracking.

Definition 4. *Dynamic topic model (DTM)*: DTM emphasizes the dynamic evolution of the existing topics and will update with the occurrence of new stories.

2.3. Basic Bayesian belief network

The basic belief network [26] derives from a probabilistic argument based on a clearly defined sample space, so it produces a network topology that separates the documents and query, which is different from the inference network.

In the basic Bayesian belief network, all index terms are interpreted as the universe of discourse U, which is taken as the sample space. Let t be the total number of terms in a collection, then $U = \{k_1, k_2, \ldots, k_t\}$, where each k_i is interpreted as an elementary concept in the space U. Further, each subset u of U is interpreted as a non-elementary concept, or a simply concept, and $g_i(u) = 1 \Leftrightarrow k_i \in u$. A document or a query is represented by a concept in the concept space U.

The probability distribution defining in the sample space U is introduced as follows, c is a concept of U, representing a document or a user query:

$$p(C) = \sum_{\forall u} p(c|u) \times p(u) \tag{1}$$

$$p(u) = \left(\frac{1}{2}\right)^t \tag{2}$$

Eq. (1) defines $p(c)$ as the degree of coverage of c in U, Eq. (2) shows that all concepts in U is equal probability.

In Fig. 1, each node d_m models a document, the node q models the user query, and k_i nodes model the terms in the collection. All of these nodes are associated with random variables denoted by the same denotations. There is an arc joining each term node k_i and each document node d_m, whenever k_i belongs to d_m. Analogously, there is an arc joining each term node k_i and query node q, whenever k_i belongs to q. In Fig. 1, the ranking computation is based on interpreting the similarity between a document d_m and the query q as an intersection of the concepts d_m and q. To quantify the degree

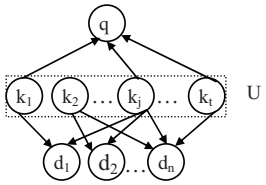

Fig. 1. Basic Bayesian belief network.

of intersection of the concept d_m, given the concept q, the probability $P(d_m|q)$ is used. Applying Bayesian theorem, the following expression can be gotten:

$$P(d_m|q) = \eta \sum_{\forall u} P(d_m|u)P(q|u)P(u) \tag{3}$$

where η is a normalizing constant. Distinct specification of conditional probabilities $P(d_m|q)$, $P(q|u)$ allow modeling different ranking strategies.

3. Topic tracking models based on Bayesian belief network model

The existing topic tracking models are divided into two categories: static topic model (STM) and dynamic topic model (DTM). In this section, using the Bayesian belief network model, we firstly propose a STM, and then give two DTM and a quantitative updating strategy.

3.1. Static topic model (BSTM)

To build BSTM, we need to extract seminal contents from existing topics identified by 1–4 samples through the method of segmentation and weight computation. Eq. (1) is the weight computation formula used in our paper [29]:

$$w(k_i) = \frac{freq(k_i) + 0.5N_{begin} + 0.5N_{end} + N_{title}}{\sum freq(k_i)} \tag{4}$$

$w(k_i)$ is the weight of term k_i, $freq(k_i)$ is the frequency of term k_i appearing in story s_i, N_{begin}, N_{end}, N_{title} mean the frequency of the term appearing in beginning, ending and title respectively, $\sum freq(k_i)$ is the summary frequency of all terms in story s_i. The top i terms sorted by weight in sample stories are used to describe topic, namely $t_j = \cup s_i = \cup\{(k_1, w_1), (k_2, w_2), \ldots, (k_i, w_i)\}$, these terms are seminal contents of topic t_j. Fig. 1 gives the topology of BSTM.

The topology of BSTM includes three level nodes: stream of new stories, term and topic, using arcs to indicate the index relationships. The contents enclosed by dashed will not change during the process of topic tracking. When a new story s_n appearing, we determine s_n whether on topic t_j through the probability computation as follows (Fig. 2):

$$P(t_j|s_n) = Sim(t_j, s_n) = \eta \sum_{\forall s} P(t_j|s) \times P(s_n|s) \times P(s) \tag{5}$$

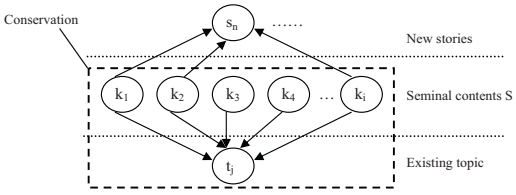

Fig. 2. BSTM based on Bayesian belief network.