Knowledge-Based Systems 89 (2015) 618–626

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Incremental learning from news events

Linmei Hu [a,*], Chao Shao [a], Juanzi Li [a], Heng Ji [b]

[a] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[b] Computer Science Department, Rensselaer Polytechnic Institute, Room 2148, Winslow, AZ, USA

ARTICLE INFO

Article history:
Received 3 March 2015
Received in revised form 4 September 2015
Accepted 7 September 2015
Available online 11 September 2015

Keywords:
News events
Topic hierarchy
Incremental learning

ABSTRACT

As news events on the same subject occur, our knowledge about the subject will accumulate and become more comprehensive. In this paper, we formally define the problem of incremental knowledge learning from similar news events on the same subject, where each event consists of a set of news articles reporting about it. The knowledge is represented by a topic hierarchy presenting topics at different levels of granularity. Though topic (hierarchy) mining from text has been researched a lot, incremental learning from similar events remains under developed. In this paper, we propose a scalable two-phase framework to incrementally learn a topic hierarchy for a subject from events on the subject as the events occur. First, we recursively construct a topic hierarchy for each event based on a novel topic model considering the named entities and entity types in news articles. Second, we incrementally merge the topic hierarchies through top-down hierarchical topic alignment. Extensive experimental results on real datasets demonstrate the effectiveness and efficiency of the proposed framework in terms of both qualitative and quantitative measures.

1. Introduction

As defined in [1,2], an event is a particular thing that happens at a certain time and location. It might be an earthquake or a presidential election. When an event happens, news articles that record its beginning, progression, impact, etc. are typically organized together, with their web page links gathered together in one special web page (e.g., the 2012 US Presidential Election[1]). In the real world, similar events on the same subject occur repeatedly. For example, the *2010 Chile earthquake*, *2011 Tokyo earthquake* and *2014 Chile earthquake* are all earthquake events irrespective of their difference in their time, locations, etc. They are likely to share a lot of common knowledge and collectively contribute to our awareness of earthquake.

In this work, inspired by the cognitive theory that people accumulate knowledge from the repeated occurrences of similar events [3], we try to automatically learn from similar events as they occur, analogous to the learning process of humans. The knowledge is represented as a topic hierarchy presenting topics at different levels of granularity. Such hierarchical topical knowledge can significantly benefit a lot of applications such as search and browsing, information retrieval and organization. For example, using such a hierarchy, one can quickly acquaint himself/herself with the comprehensive knowledge about the subject of the events (e.g., earthquake knowledge) accumulated from real events (e.g., earthquake events). It not only provides a structured topical summarization of the events but also serves as prior knowledge to improve document organization and topic extraction for new similar events [4,5].

Though a lot of research has been conducted on events [6–8] and topic hierarchy construction from events [9–15], it remains unknown how to incrementally learn from similar news events in order to acquire a comprehensive topic hierarchy representing the knowledge about the subject of the events. An obvious approach is to directly apply existing topic hierarchy construction methods such as hierarchical topic models on the news articles of all the events. However, such a method is unjustifiable for three reasons. First, the data will be too big to model together, which will lead to extremely low efficiency. Second, the data of different events has big differences in the contents and named entities. Thus, it is not proper to mix the data for modeling together, as we will see in experiments. Last, when a new event comes, the method needs to model all the data from scratch, which is inflexible and time-consuming.

Therefore, in this work, we propose a simple two-phrase framework to address the problem. In the first phase, similar to [16], we present an efficient algorithm to recursively construct topic hierarchies for each event based on a novel topic model EETM (Event

* Corresponding author. Tel.: +86 (010) 62789831; fax: +86 (010) 62781461.
E-mail addresses: hulinmei1991@gmail.com (L. Hu), shaochao@keg.cs.tsinghua.edu.cn (C. Shao), ljz@keg.cs.tsinghua.edu.cn (J. Li), jih@rpi.edu (H. Ji).

[1] http://www.reuters.com/news/archive/GCA-Elections2012.

http://dx.doi.org/10.1016/j.knosys.2015.09.007
0950-7051/

Entity Topic Model). The model models a topic as three distributions: one over words, one over named entities and one over named entity types. In this way, when aligning topics in different events, we can easily circumvent the named entities to find matching topics using only its word distributions. In addition, the distribution of topics over named entity types shows which types of entities a topic tends to and thus can benefit topic visualization. We do not use the existing hierarchical topic models to directly construct the event topic hierarchy due to their computational inefficiency. In the second phase, we propose the algorithm IHTA (Incremental Hierarchical Topic Alignment) to incrementally integrate the event topic hierarchies in the first phase into a comprehensive topic hierarchy for the subject of the events by aligning the new event topic hierarchy with the target topic hierarchy in a top-down fashion. This algorithm is flexible and can handle new coming events easily.

The main contributions of this paper can be summarized as follows:

(1) We propose and formally define the problem of incremental learning from news events on the same subject, analogous to the process of human learning.
(2) We present an efficient two-phase framework to address the proposed problem effectively. In the first phase, we propose a novel event entity topic model (EETM) to mine topics from news events considering named entities and their types, based on which we recursively construct topic hierarchies for the events. In the second phase, we present an incremental hierarchical topic alignment (IHTA) algorithm to incrementally learn the topic hierarchy for the subject of the events.
(3) We extensively evaluate our two-phase framework with real datasets in two languages. Experimental results demonstrate the effectiveness and efficiency of the incremental framework in terms of both qualitative and quantitative measures.

The rest of the paper is organized as follows. In Section 2, we will formalize the problem. In Section 3, we will illustrate the two-phase framework. We will provide empirical evidence for the effectiveness of the proposed framework in Section 4. We will review related work in Section 5. Finally, we will summarize the proposed framework and discuss future research directions in Section 6.

2. Problem definition

In this section, we introduce some concepts and formally define the task of learning a topic hierarchy from news events incrementally.

News event. An event (or news event) e is defined as a particular thing that happens at a certain time and place [1,2]. Each event consists of a collection of news articles $D_e = \{d_1, d_2, \ldots, d_M\}$. A document d_i is a sequence of words $\{w\}$ and named entities $\{e\}$ with their types $\{\tau\}$ (generated by parts of speech).

We take events on the same subject as similar events. For example, various earthquake events (e.g., *2010 Chile earthquake* and *2011 Tokyo earthquake*) are all about the subject of earthquake, thus are considered as similar events irrespective of their difference in the named entities (e.g., time, locations, organizations, persons).

Topic. A topic of an event is one aspect covered by the news documents of the event. Formally, a topic t is defined as three distributions, the distribution over words $\beta = P(w|t)$ ($w \in V$, where V denotes the word vocabulary set), the distribution over entities $\tilde{\beta} = P(e|t)$ ($e \in \tilde{V}$, where $\tilde{V}$ is the entity vocabulary set) and the distribution over entity types $\psi = P(\tau|t)$ ($\tau \in V_\tau$, where V_τ is the vocabulary set of entity types). Then, a topic can be represented as $t = (\beta, \tilde{\beta}, \psi)$.

Intuitively, a topic of an event is semantically coherent in the sense that the high probability words, named entities and entity types collectively suggest a theme. For example, the topic "*Tsunami*" of the "*2010 Chile earthquake*" may assign higher probabilities to words such as "*ocean*", "*disaster*", and "*wave*", named entities such as "*Chile*", "*Peru*", and "*the Pacific Ocean*", and entity types such as organizations and locations. News documents are divided into groups according to their topics. Each topic corresponds to a group of news documents, denoted by $D_t = \{d\}$.

Topic hierarchy. A topic hierarchy is defined as a tree which consists of a set of topics and the links between topics. We denote the set of all the topics including the root topic in a hierarchy as $T = \{t_1, t_2, \ldots t_{|T|}\}$. The set of all links between topics and subtopics is represented as $R = \{R_j | R_j = (t_i, t_j, p_{t_j})\}$ where $t_i, t_j \in T$, t_j is a subtopic of t_i and p_{t_j} is the probability of generating the link l_j from t_i to t_j. The sum of the probabilities of the links pointing from the same super-topic sums to 1.0. Finally, we denote a topic hierarchy as $H = (T, R)$.

In our work, there are two kinds of topic hierarchies: event topic hierarchy for a single event and subject topic hierarchy as the learnt knowledge about the subject of the similar events. The latter is the focus of this paper.

The task of incremental learning from news events. Given a set of similar events $E = \{e_1, e_2, \ldots e_n\}$ associated with their news document collections $C = \{D_1, D_2, \ldots D_n\}$, our goal is to automatically learn a comprehensive subject topic hierarchy $H = (T, R)$ which represents the knowledge of the subject of the events.

Though the events are similar, it is not proper to model them together due to huge difference in their contents and named entities, as we will see in experiments. Therefore, we model a topic hierarchy $H_i = (T_i, R_i)$ for each event $e_i \in E$ and then incrementally integrate them via top-down hierarchical topic alignment to get the subject topic hierarchy H. The detail is in the following section.

3. Our two-phase framework

In this section, we present our flexible two-phase framework for incremental learning from news events. It includes two phases:

1. Event topic hierarchy construction. For each event, we recursively construct a topic hierarchy from its news documents based on a new topic model EETM. Every time when we apply EETM, the documents are divided into clusters according to their most probable topics. EETM is applied recursively on the clusters, so that the documents are divided into sub-clusters according to the subtopics. In this way, an event topic hierarchy can be formulated.

2. Incremental hierarchical topic alignment. Based on the event topic hierarchies constructed in the first phase, we learn a comprehensive subject topic hierarchy for the subject of similar events through IHTA. Every time a new event comes, we update the learnt subject topic hierarchy by aligning it with the new event topic hierarchy. In this way, we not only avoid mixing up events but also deal with new coming events easily.

We will introduce the two phases in detail in the following sections.

3.1. Event topic hierarchy construction

Inspired by [11], we apply a recursive algorithm based on our proposed EETM to generate a topic hierarchy for a news event in a top-down fashion. The algorithm is as follows: