Front. Comput. Sci., 2016, 10(1): 136–146
DOI 10.1007/s11704-015-4548-5

RESEARCH ARTICLE

Topic hierarchy construction from heterogeneous evidence

Han XUE[1,2], Bing QIN[1], Ting LIU (✉)[1], Shen LIU[1]

1 School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
2 Harbin Engineering University Library, Harbin Engineering University, Harbin 150001, China

Abstract Existing studies on hierarchy construction mainly focus on text corpora and indiscriminately mix numerous topics, thus increasing the possibility of knowledge acquisition bottlenecks and misconceptions. To address these problems and provide a comprehensive and in-depth representation of domain specific topics, we propose a novel topic hierarchy construction method with real-time update. This method combines heterogeneous evidence from multiple sources including folksonomy and encyclopedia, separately in both initial topic hierarchy construction and topic hierarchy improvement. Results of comprehensive experiments indicate that the proposed method significantly outperforms state-of-the-art methods (t-test, p-value < 0.000 1); *recall* has particularly improved by 20.4% to 38.7%.

Keywords hierarchy construction, Chinese topic hierarchy, folksonomy, heterogeneous evidence, hierarchy update

1 Introduction

Topic hierarchy is a fine-grained hierarchy that can be established for various topics and can thus provide a comprehensive, in-depth, and up-to-date picture of domain specific topics. Figure 1 shows an excerpt of a sample topic hierarchy in the movie domain on the root topic 励志 "Encouragement". Nodes with gray backgrounds represent the movie resources, and nodes with white backgrounds represent social tags. Automatic topic hierarchy construction is an important task in the field of natural language processing, knowledge management, and semantic web. This task can promote the development of related tasks, such as information retrieval and navigation, question answering, and recommendation system.

Traditional hierarchies are mainly generated manually or semi-automatically from text corpora by a few experts. This process is time consuming and difficult to update without public participation. Moreover, finding a text corpus that can accurately describe a highly specialized or ever-changing topic is complicated [1]. If available, then going through the entire text corpus and catching up with all newly emerging topics are impossible tasks for humans. For example, obtaining a text corpus consisting of a formal description about an uncommon topic, such as "Cult", is challenging. However, tags provide a flexible approach that can be used to characterize topic "Cult", including cult, non-mainstream, and small budget. Therefore, a few researchers propose the use of folksonomy [2] instead of text corpus. Folksonomy emerges when users with diverse expertise are authorized to freely annotate resources of interest with arbitrary words (i.e., tags) and then share these tags with one another. These tags have rich semantic content and real-time update. Folksonomy can provide an effective method that can be used to promote the development of a traditional hierarchy construction.

However, previous methods mainly construct a single hierarchy of multiple topics without being distinguished by different topics. These methods may lead to misconception because the common terms in various topics have different semantics. 爱情 "Love" is a key topic in the movie domain, but this term is also a sub-topic of 励志 "Encouragement" (Fig. 1). Defining a term with completely different meanings in the same hierarchy is inaccurate. Ignoring a meaning of a term is also incorrect. Thus, we study topic hierarchy construction

Received November 30, 2014; accepted May 19, 2015

E-mail: tliu@ir.hit.edu.cn

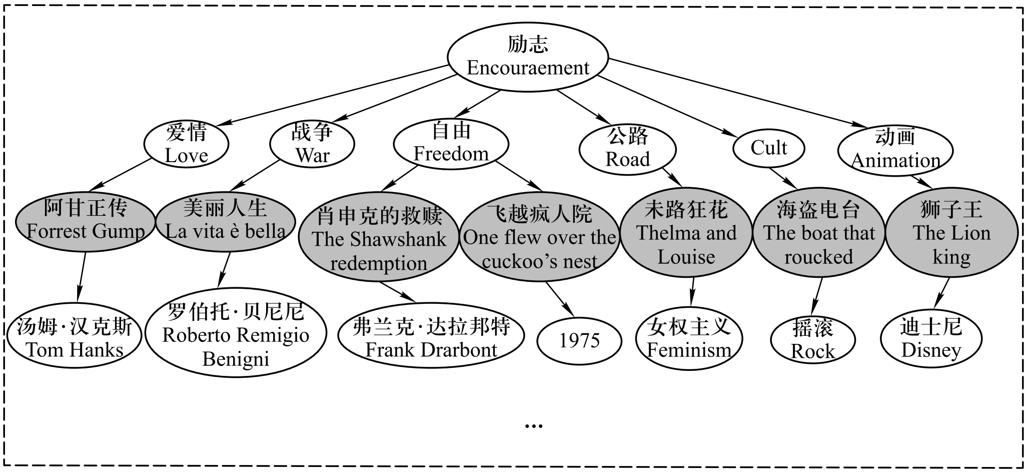

Fig. 1 Excerpt of one sample topic hierarchy

from folksonomy to address the misconception.

Although folksonomy provides an easy way to obtain domain knowledge from the public, folksonomy itself does not provide explicit relations, which can be obtained from domain-independent encyclopedia knowledge sources. For example, if we want to know the public opinion for a movie, we will refer to Douban.com Movie[1] and Baidu Video[2]. Both Douban.com Movie and Baidu Video are popular social media sites that support folksonomy in China. However, Chinese Wikipedia[3] is a better choice when we require authoritative knowledge about a movie. Furthermore, information from multiple sources provides clues in different views [3,4] and helps overcome the bias of any single source.

On the basis of this background, we propose an automatic topic hierarchy construction from Chinese heterogeneous evidence. The task is composed of topic term extraction, topic relation identification, and topic hierarchy construction. Given a collection of Chinese information sources about a film, including Douban.com Movie, Baidu Video, and Chinese Wikipedia, we first identify the topics from these sources. The terms are then ranked on the basis of the importance scores for a certain topic to determine the root topics and candidate sub-topics. Considering the characteristics of information sources, we leverage heterogeneous evidence from multiple sources for topic relation identification. For example, we design local and global semantic similarities as undirected evidence for implicit topic relation determination, whereas directed evidence such as category and infobox from Chinese Wikipedia can be used for explicit topic relations extraction. Finally, we propose a novel two-step combination scheme of directed and undirected evidence for initial topic hierarchy generation and topic hierarchy improvement. The contributions can be summarized as follows:

- We present an automatic topic hierarchy construction method from scratch with real-time update. This method can provide a comprehensive, in-depth, and up-to-date picture of domain specific topics.
- We extract heterogeneous evidence from multiple sources by exploring their unique characteristics and adopt a novel combination scheme by leveraging the strength of undirected and directed evidence.
- We are the first to learn topic hierarchy construction from Chinese folksonomy in the field of Chinese ontology construction. Moreover, our proposed method is unsupervised and language independent. Hence, our method is applicable to enormous information and other languages.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 presents the proposed method. Section 4 discusses our evaluation and results. Finally, Section 5 concludes this paper.

2 Related work

Existing methods for hierarchy construction are generally based on either text corpora or folksonomy. In this section, we briefly introduce the two types of related works.

[1] http://movie.douban.com
[2] http://video.baidu.com
[3] http://zh.wikipedia.org/wiki/

Considerable research has been conducted on text-based hierarchy construction. Pattern-based methods have been applied to extract various types of semantic relations, including is-a relations [5], part-of relations [6], and so on. However, pattern-based methods suffer from the sparse coverage of patterns in a given corpus. Clustering-based methods allow the discovery of relations, which inexplicitly appear in the text, but fail to produce coherent and accurate results compared with pattern-based methods. Ming et al. [7] present a prototype hierarchy-based clustering framework for organizing web collections. However, this framework has a limited capability to create new categories in an existing hierarchy established by experts. By taking advantage of pattern- and clustering-based methods, Snow et al. [8] introduce a probabilistic model to determine the most possible hierarchy for a set of terms. This model attaches new terms under the appropriate nodes of an existing hierarchy. Yang and Callan [9] propose a metric-based framework that integrates various features by minimizing the change of information functions for automatic hierarchy induction. Yu et al. [10] extend this method by employing more objective functions, including minimum hierarchy discrepancy both from local and global aspects and minimum semantic inconsistency. Navigli et al. [11] train classifiers to detect is-a relations between terms and weight the link with the number of traversed nodes to extract an optimal hierarchy from the resultant hypernymy graph. By contrast, Zhu et al. [4] consider a graph-based method with estimated weight instead of simple 0/1 counts to incrementally generate a hierarchy by using several pieces of evidence from a given collection of user-generated content (UGC). UGC includes blogs, cQAs, and tweets on a specific topic, as well as external knowledge from Wikipedia, WordNet, and search engine results. Hoffart et al. [3] propose the integration of information from Wikipedia, WordNet, Geo-Name corpus, etc., to develop Yago2, which is an open domain-structured hierarchy.

Folksonomy can overcome the knowledge acquisition bottleneck better than traditional text corpora. The tagging-and-sharing process generates potentially valuable semantic information. However, studies on hierarchy construction from folksonomy have recently begun and only a few mature text-based methods have adjusted to folksonomy based on the the tag itself. An unsupervised clustering model for deterministic annealing [12] is used to derive hierarchies from a set of tags on the basis of co-occurrence information. Heymann and Garcia-Molina [13] propose a hierarchical-clustering method based on greedy algorithm for hierarchy generation by using graph centrality in a similarity graph based on the co-occurrence of tags. Angeletou et al. [14] map tags to WordNet and extract synonyms and hypernyms for tag set extension. By connecting the extended tag set and semantic entities of Watson, a search engine of semantic web, they determine the relations between pair-wise tags and construct a hierarchy called FLOR. However, a large number of tags cannot be mapped to WordNet because of their coverage and randomness. Tomuro and Shepitsen [15] use Wikipedia as the external knowledge source for tag domains and apply a hierarchical agglomerative clustering algorithm to develop a hierarchy of tags. Liu et al. [1] use a general-purpose knowledge base called Probase, as well as keyword search by a commercial search engine, to supply the required knowledge and context for a set of keyword phrases. They also develop a Bayesian method to derive a hierarchy from the set of keywords. Several researchers introduce the concept of latent dirichlet allocation (LDA) [16] for the hierarchical construction of tags, such as Tag-Topic approach [17], User-Word-Topic approach [18], and Actor-Concept-Instance-Topic approach [19].

In summary, most of the aforementioned methods mainly construct hierarchy without being distinguished by different topics. Such scenario may cause misconceptions. Therefore, more topic hierarchies should be considered instead of merely a single hierarchy of undistinguished multiple topics. Other text-based methods that consider tag characteristics should also be used in folksonomy.

3 Method

In this section, we describe the proposed method in detail. We first introduce the information source set (Fig. 2) and then present the three sub-tasks of the proposed method.

3.1 Information sources

To construct high-quality hierarchies, we consider multiple information sources that not only include folksonomies in specific domains, but also domain-independent encyclopedia sources. In the case of film, which is a fast-changing domain that is close to the real life of people, three typical Chinese information sources are available:

- **Douban.com Movie** is a famous Chinese folksonomy source that can provide timely and large amounts of user-generated tags to specific movies. However, these tags are only organized by frequency and have varying qualities because of randomness and ambiguity.
- **Baidu Video** is another famous Chinese folksonomy

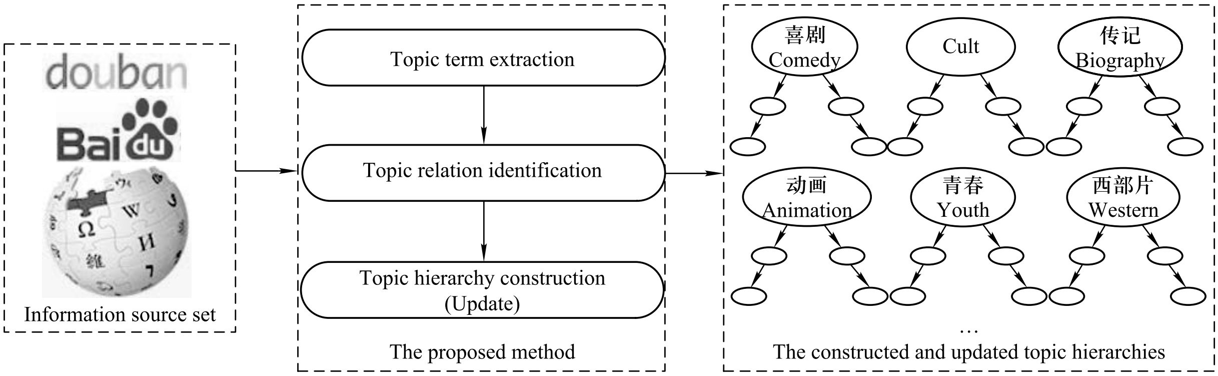

Fig. 2 Workflow of the proposed topic hierarchy construction from heterogeneous evidence

source related to film and has well-written tags. This source has relatively few tags and requires human editing before being published.

- **Chinese Wikipedia** is a domain-independent encyclopedia source that depends on humans for compiling and updating. Although sufficiently organized into structured formats that can be easily accessed, this source cannot keep up with the ever-changing Internet.

To overcome the bias of any single information source, we combine the power of two prevailing folksonomy information sources, namely, Douban.com Movie and Baidu Video to obtain knowledge from the public. Moreover, we leverage the knowledge of professionals by using Chinese Wikipedia.

3.2 Problem formulation

Preliminary 1 A topic hierarchy is defined as a tree that consists of a set of unique terms specific to the topic and a set of relations between these terms.

We define the terms as follows:

- **Information source set** $S_y = \{S_{\text{Douban.com Movie}}, S_{\text{Baidu Video}}, S_{\text{Chinese Wikipedia}}\}$;
- **Resource set** $S = \{s_1, s_2, \ldots, s_i, \ldots\}$ denotes a set of resources of interest, such as movies, where s_i indicates a document that consists of terms relative to the resource. For example, a movie 罗马假日 “Roman Holiday” can be represented as a document consisting of tags relative to this movie, such as 浪漫 “Romance”, 约会 “Dating”, and 奥黛丽· 赫本 “Audrey Hepburn”.
- **Topic set** $Z = \{z_1, z_2, \ldots, z_i, \ldots\}$ denotes a set of potential topics that exists in the resource set, where z_i indicates one of the topics.

For a specific topic z_i, we define a topic hierarchy as $H = \{T, R\}$, which includes the following terms:

- **Topic term set** $T = \{w_1, w_2, \ldots, w_i, \ldots\}$ denotes a set of terms extracted from resource set S specific to the topic, where w_i indicates a topic term. Topic term set T includes root topic C and potential sub-topics of C.
- **Topic relation set** $R = \{r_1, r_2, \ldots, r_i, \ldots\}$ denotes a set of relations between the topic terms in T, where $r(w_i, w_j)$ indicates a directed link from w_i to w_j. This set links all the terms in T into a hierarchy rooted at C.

Given a collection of information sources, we first collect information from S_y and organize to resource set S. We then identify potential topics Z from resource set S. For a specific topic, we extract relative topic terms T and identify topic relations R. A topic hierarchy H is then generated on the basis of the relevant topic terms and topic relations between them. The following sections describe the proposed method in detail.

3.3 Topic term extraction

Terms are the building blocks of a hierarchy. A single hierarchy of multiple topics may result in misconception. Therefore, we propose a topic term extraction to address the misconceptions. Thus, we extract and organize terms on the basis of the topics learned from the corpus. As candidate terms, tags are extracted by public participation. This condition avoids a series of natural language processing tasks, namely, word segmentation and syntactic parsing. However, topic term extraction is complicated because of the considerable variations in tag quality owing to randomness and ambiguity. Hence, we have to first identify the topics from a resource set and determine the topic distribution of the terms through LDA [16]. The terms are then ranked based on the importance scores $R_z(w)$ for a certain topic through topic-sensitive random walk [20]. The top-ranking term that best represents the corresponding topic is extracted as root topic C, whereas other relevant terms are extracted as candidate sub-topics of C. Both root topic and candidate sub-topics constitute the topic term set.

3.4 Topic relation identification

Topic relation identification is an essential subtask of topic hierarchy construction after topic term extraction. To infer the topic relation $r(w_i, w_j)$, we need to estimate the probability $p(r(w_i,w_j))$ that the term w_j is a sub-topic ofw_i. Take the topic 成长 "Growth" as an example. If we know the probability $P(r(\text{Growth, Campus}))$ that the term 校园 "Campus" is a sub-topic of 成长 "Growth" is higher than a few thresholds, a condition that indicates that a relationship between the two terms r (Growth, Campus) exists, and then a path is probable between the two terms on the topic hierarchy.

Influenced by Ref. [9], we estimate the probability $p(r(w_i, w_j))$ by using heterogeneous evidence from multiple sources. We disregard pattern-based evidence, such as hypernym [5], meronym [6], search engine [4], and WordNet similarity [4], because patterns have low coverage on folksonomy tags. Considering the association characteristics of folksonomy, we use both local and global similarities as evidence because our preliminary experiments indicate that this method can connect the most related terms under a specific topic, as explained by Xue et al. [20]. We also use the infobox, category, and redirect from Chinese Wikipedia. The details of the evidence are listed as follows:

- $e_{\text{local}}(w_i, w_j)$ denotes the local semantic similarity between terms w_i and w_j. This evidence can be estimated as the ratio of the number of co-occurrences of the two terms in the same resource assigned to a certain topic and the number of co-occurrences of the two terms in the same resource.

$$e_{\text{local}}(w_i, w_j) = \frac{C^S_{w_i,w_j,z}}{C^S_{w_i,w_j}}. \tag{1}$$

- $e_{\text{global}}(w_i, w_j)$ denotes the global semantic similarity between terms w_i and w_j. This evidence is defined as the cosine similarity of the two terms over all the topic dimensions considered in the entire resource collection S.

$$e_{\text{global}}(w_i, w_j) = \frac{\sum_z^S p(z|w_i)p(z|w_j)}{\sqrt{\sum_z^S p(z|w_i)^2 \sum_z^S p(z|w_j)^2}}. \tag{2}$$

The topic distribution $p(z|w_i)$ of each term is computed through LDA as follows:

$$p(z|w_i) = \frac{p(z)p(w_i|z)}{\sum_{z'} p(z')p(w_i|z')}. \tag{3}$$

- $e_{\text{source}}(w_i, w_j)$ returns "1" when terms w_j and w_i occur in the same resource about a movie titled w_i and "0" if otherwise.
- $e_{\text{redirect}}(w_i, w_j)$ returns "1" when terms w_j and w_i can be redirected to each other in Chinese Wikipedia and "0" if otherwise.
- $e_{\text{category}}(w_i, w_j)$ returns "1" when page w_j in the Chinese Wikipedia has category w_i and "0" if otherwise.
- $e_{\text{infobox}}(w_i, w_j)$ returns "1" when the infobox of page w_j in Chinese Wikipedia has term w_i and "0" if otherwise.

Contrary to simply adding or multiplying the evidence [9,10], we divide them into directed and undirected evidence according to Zhu et al. [4]. Table 1 shows that all directed evidence include evidence that can be used for explicit topic relations extraction, whereas all undirected evidence includes evidence that can be used for implicit topic relations discovery between topic terms. Then, we propose a two-step scheme to combine the directed and undirected evidence to infer the topic relation. First, we use a linear combination to estimate the probability $p(r(w_i, w_j))$ by using Eq. (4), wherein Eqs. (1) and (2) are substituted.

$$p(r(w_i, w_j)) = e_{\text{local}}(w_i, w_j)((1-\rho)e_{\text{global}}(w_i, w_j) + \rho). \tag{4}$$

On the basis of $p(r(w_i, w_j))$, we form the candidate topic relation set for topic hierarchy construction. These pieces of heterogeneous evidence contribute both in determining the topic relation and in constructing the topic hierarchy.

Table 1 Sources of evidence

		Source
Directed evidence	e_{source}	S_y
	e_{category}	$S_{\text{Chinese Wikipedia}}$
	e_{infobox}	$S_{\text{Chinese Wikipedia}}$
Undirected evidence	e_{local}	S_y
	e_{global}	S_y
	e_{redirect}	$S_{\text{Chinese Wikipedia}}$

3.5 Topic hierarchy construction

Topic hierarchy construction is the core of our work and can be divided into initial topic hierarchy generation (Algorithm 1: lines 1 to 18) and topic hierarchy improvement (Algorithm 1: lines 19 to 52). Given a specific topic z, topic term set T, and original topic relation set R, we use an iterative graph-based method to generate an initial topic hierarchy. We only add one topic term into the hierarchy at each step, thus our approach is amendable for the incremental updating of the initial topic hierarchy. Algorithm 1 shows that S stands for the resource set that consists of all terms from information source set S_y, whereas T denotes the candidate topic term set, including the root topic C and potential sub-topics of C. In the ith iteration, we add a topic term w in $T{-}T_{i-1}$ into T_{i-1} by using Eq. (5) to maximize the overall relatedness between w and all topic terms in T_{i-1} (lines 3 to 9).

$$w = \mathop{\arg\max}_{w_s \in T - T_{i-1}} \sum_{w_k \in T_{i-1}} p(r(w_k, w_s)). \tag{5}$$

After adding w into T_{i-1}, the edges between w and the topic terms in T_{i-1} are added into R_{i-1} (lines 10 to 15). To distinguish the importance of topic relations in generating the topic hierarchy, we weigh each edge in R_i (line 16) with $q(w_s \rightarrow w_k)$ by using Eq. (6). We denote $L = \{w_u \rightarrow w_{u+1}\}_{u=0}^{|L|-1}$ as a path ends with root topic and w_k. The score of L is calculated as Eq. (7), where $R_z(w_u)$ is the important score of the term w_u under the specific topic z, as described in Section 3.3. We need to determine an optimal subset of R_i (line 17) by applying Chu-Liu/Edmonds' optimum branching algorithm [21].

$$q(w_s \rightarrow w_k) = \max_{L: w_s \rightarrow w_k} score_L. \tag{6}$$

$$score_L = \sum_{u=0}^{|L|-1} R_z(w_u) p(r(w_u, w_{u+1})). \tag{7}$$

Algorithm 1   Topic hierarchy construction algorithm

```
Input:
    S: the resource set that consists of all the terms
    T: the candidate topic term set
Output:
    T_ret: the topic term set of the resultant hierarchy
    R_ret: the topic relation set of the resultant hierarchy
 1: Initialize $T_0 = \{C\}, R_0 = \emptyset$
 2: for i = 1 TO $\infty$ do
 3:     $w \leftarrow$ selectTermFrom($T - T_{i-1}$)
 4:     if $w$ is NIL then
 5:         $R_{ret} \leftarrow R_{i-1}$
 6:         $T_{ret} \leftarrow T_{i-1}$
 7:         break
 8:     end if
 9:     $T_i \leftarrow w \cup T_{i-1}$
10:     $R_i \leftarrow R_{i-1}$
11:     for $w_k$ IN $T_{i-1}$ do
12:         if edgeBetween($w_k, w$) exists then
13:             $R_i \leftarrow R_i \cup$ edgeBetween($w_k, w$)
14:         end if
15:     end for
16:     edgeWeighting($R_i$)
17:     $R_i \leftarrow$ hierarhcyPruning($R_i$)
18: end for
19: for $w$ IN $T_{ret}$ do
20:     if $\exists w_f \in S, w_f \neq w$ and $e_{source}(w_f, w) = 1$ then
21:         $T_{ret} \leftarrow T_{ret} \cup w_f$
22:         $R_{ret} \leftarrow R_{ret} \cup$ edgeBetween($w_f, w$)
23:     end if
24: end for
25: for $w$ IN $T_{ret}$ do
26:     $T_c = \emptyset$
27:     for $w_c$ IN $S$ do
28:         if $e_{category}(w_c, w) = 1$ then
29:             $T_c \leftarrow T_c \cup w_c$
30:         end if
31:     end for
32:     $t \leftarrow$ selectCategoryFrom($T_c$)
33:     $T_{ret} \leftarrow T_{ret} \cup t$
34:     $R_{ret} \leftarrow R_{ret} \cup$ edgeBetween($t, w$)
35:     $T_{in} = \emptyset, R_{in} = \emptyset$
36:     for $w_{in}$ IN $S$ do
37:         if $e_{infobox}(w, w_{in}) = 1$ then
38:             $T_{in} \leftarrow T_{in} \cup w_{in}$
39:             $R_{in} \leftarrow R_{in} \cup$ edgeBetween($w, w_{in}$)
40:         end if
41:     end for
42:     $T_{ret} \leftarrow T_{ret} \cup T_{in}$
43:     $R_{ret} \leftarrow R_{ret} \cup R_{in}$
44: end for
45: for $w_a$ IN $T_{ret}$ do
46:     for $w_b$ IN $T_{ret}$ do
47:         if $w_a \neq w_b$ and $\mathrm{e}_{redirect}(w_a, w_b) = 1$ then
48:             $R_{ret} \leftarrow$ mergeedgeBetween($w_a, w_b$)
49:             $T_{ret} \leftarrow T_{ret} - w_b$
50:         end if
51:     end for
52: end for
```

To guarantee that the topic hierarchy can provide a comprehensive and in-depth picture of the topic, we further improve the initial topic hierarchy by using directed heterogeneous evidence. We import the missing film name w_f for w in T_{ret} when $e_{\text{source}}(w_f, w) = 1$ (lines 19 to 24). If film w in T_{ret} has many categories ($e_{\text{category}}(w_c, w) = 1$), we select the most related category as the parent node through Eq. (8) (lines 25 to 34) because excessive categories may lead to false associations with other movies. In Eq. (8), T_c denotes the category set and C is the root topic of T_{ret}.

$$t = \mathop{\arg\max}_{t_c \in T_c, w \in T_{ret}} (p(r(w, t_c) + p(r(t_c, C)). \tag{8}$$

We then introduce explicit properties about the film with structured infobox T_{in} (lines 35 to 44) as the child of term w. Finally, we merge the film name and alias (lines 45 to 52) through undirected evidence e_{redirect} and adjust the nodes and edges accordingly. The resultant topic hierarchy is $H = \{T_{\text{ret}}, R_{\text{ret}}\}$.

3.6 Topic hierarchy update

To keep up with the rapid-changing Internet, the proposed topic hierarchy construction algorithm can be used to incrementally update the previous topic hierarchy with the newly obtained data. We let $H_{\text{old}} = \{T_{\text{old}}, R_{\text{old}}\}$ be the existing topic hierarchy, S_{new} is the new resource set that contains previous and newly obtained data ($S = S_{\text{new}}$), and T_{add} is the new emerging topic term identified from S_{new}. When a root topic identified from S_{new} is not in the topic set Z, we initialize T_0 with the new root topic. We use Algorithm 1 to generate a new topic hierarchy $H_{\text{new}} = \{T_{\text{new}}, R_{\text{new}}\}$, wherein $R_0 = \emptyset$; Otherwise, we initialize $T_0 = T_{\text{old}}$, $R_0 = R_{\text{old}}$, and $T = T_{\text{add}}$. Thereafter, we use Algorithm 1 to add new nodes in T_{add} into T_{old} and the edges between topic terms in T_{add} and T_{old} into R_{old}. The update process can find new emerging root topics and construct the corresponding topic hierarchy. Moreover, this process can break the existing relation between two nodes and create a better hierarchy instead of merely adding the new topic term as a child of any one of the node. The reason behind these achievements is that the proposed method is based on the calculation of heterogeneous evidence and not only the change of the structure.

4 Evaluation

4.1 Data and experimental setup

To provide a comprehensive and in-depth picture of a specific domain, we collect data from multiple film sources and construct a few topic hierarchies on the basis of film topics. The method is domain independent and can be applied to any other domain. The dataset contains the top 250 movies from Douban.com Movie. $S_{\text{Douban.com Movie}}$ is the tag used by the public to refer to the top 250 movies crawled from Douban.com through the Douban.com API. For $S_{\text{Baidu Video}}$, we obtain the tags of the top 250 movies by submitting the movie name as the query on the Baidu Video website. $S_{\text{Chinese Wikipedia}}$, which is composed of infobox, category, and redirect about the top 250 movies, is crawled from Chinese Wikipedia. The data in all three sources are divided into original and updated sets on the basis of two different crawl times about the top 250 movies. The first is crawled on June 2012 and the second is up to May 2014. The latter has 43 movies that are different from the former. Table 2 shows the concise statistics of the dataset.

After analyzing the models with different numbers of topics ranging from 10 to 100 according to the size of the dataset, we set the number of initial topics to 40, which provides the best performance. Then, we run LDA with 1 000 iterations of Gibbs sampling. We terminate the algorithm of topic-sensitive random walk [20] in the topic term extraction when the number of iterations reaches 100 or when the difference in the importance scores for each term between two neighbor iterations is less than 0.000 001. The grid-search algorithm is also applied to the parameter selection.

Table 2 Statistics about crawled data

		Number of terms
Original 250 movies	$S_{\text{Douban.com Movie}}$	2 459
	$S_{\text{Baidu Video}}$	967
	$S_{\text{Chinese Wikipedia}}$	5 956
Updated 43 movies	$S_{\text{Douban.com Movie}}$	517
	$S_{\text{Baidu Video}}$	120
	$S_{\text{Chinese Wikipedia}}$	1 188

4.2 Evaluation metrics

To assess the quality of topic hierarchy, we perform an automatic evaluation against a manually created gold standard. For each topic, two annotators are employed to create topic hierarchies independently by using candidate topic terms following three rules.

Rule 1: Relevancy

All nodes on the hierarchy should be reasonable sub-topics of the root topic to guarantee that noisy information will be excluded.

Rule 2: Coverage

All relevant nodes should be included into the topic hierarchy to ensure that any useful information about the root topic will be included.

Rule 3: Structure

Each two connected nodes should be directly related such that no other nodes in the candidate topic set can be inserted between them. This approach will make the topic hierarchy completely structured.

For example, given the candidate topic term set {灾难 “Disaster”, 票房 “Box Office”, 1997, 泰坦尼克号 “Titanic”, 对她说 “Talk to her”, 女性主义 “Feminism”, 西班牙 “Spain”} for root topic 爱情 “Love”. First, only 票房 “Box Office” will be removed on the basis of Rules 1 and

2. According to Rule 3, the gold standard hierarchy should be {爱情 “Love” → 灾难 “Disaster”, 灾难 “Disaster” → 泰坦尼克号 “Titanic”, 泰坦尼克号 “Titanic” → 1997, 灾难 “Disaster” → 女性主义 “Feminism”, 女性主义 “Feminism” → 对她说 “Talk to her”, 对她说 “Talk to her” → 西班牙 “Spain”}. The two annotators compare their candidate hierarchies and establish the gold standards through discussions (Kappa 0.92). The evaluation metrics are precision, recall, and F_1-score (F_1). We denote R_{method} and R_{gold} as the topic relation set generated by a few methods that need to be evaluated and the gold standard, respectively. The evaluation metrics are represented as follows:

$$\begin{aligned} precision &= \frac{R_{\text{method}} \cap R_{\text{gold}}}{R_{\text{method}}}, \\ recall &= \frac{R_{\text{method}} \cap R_{\text{gold}}}{R_{\text{gold}}}, \\ F_1 &= 2 \cdot \frac{precision \cdot recall}{precision + recall}. \end{aligned} \tag{9}$$

4.3 Ablation study on information sources

We conduct an ablation study on different information source sets to analyze the effects of each information source on the topic hierarchy construction. In the implementation, the information source sets contain 1) $S_{\text{Douban.com Movie}} + S_{\text{Baidu Video}}$ (no Chinese Wikipedia), 2) $S_{\text{Douban.com Movie}} + S_{\text{Chinese Wikipedia}}$ (no Baidu Video), 3) $S_{\text{Baidu Video}} + S_{\text{Chinese Wikipedia}}$ (no Douban.com Movie), and 4) $S_{\text{Douban.com Movie}} + S_{\text{Baidu Video}} + S_{\text{Chinese Wikipedia}}$ (All). Table 3 reveals the need to consider different information sources. The combination of data from Douban.com Movie, Baidu Video, and Chinese Wikipedia performs significantly better than the other three combinations by 9.6% to 42.4% on the average F_1-score (t-test, p-value < 0.000 1). Three reasons are cited for this result. First, more information sources provide more topic terms and the relations for the recall is significantly improved by 11.6% to 41.8% (t-test, p-value < 0.000 1). Second, a set of information from more sources provides more semantic evidence that can help achieve better results by overcoming the bias of any single information source. The precision is significantly improved by 8.1% to 41.3% (t-test, p-value < 0.000 1). Third, more types of information sources provide more kinds of semantic evidence. We can leverage the strengths of heterogeneous semantic evidence. Domain-dependent information sources, such as Douban.com Movie and Baidu Video, provide undirected evidence that can help determine the implicit topic relations to generate the initial topic hierarchy. By contrast, domain-independent Chinese Wikipedia provides directed evidence that can help determine the explicit topic relations to improve the initial topic hierarchy. When comparing the other three results, $S_{\text{Douban.com Movie}} + S_{\text{Chinese Wikipedia}}$ (no Baidu Video) performs better than $S_{\text{Baidu Video}} + S_{\text{Chinese Wikipedia}}$ (no Douban.com Movie) because Douban.com Movie provides timely user-generated knowledge that is useful for topic hierarchy construction. $S_{\text{Douban.com Movie}} + S_{\text{Baidu Video}}$ (no Chinese Wikipedia) performs worst because we cannot obtain directed evidence for explicit topic relation discovery and topic hierarchy improvement without Chinese Wikipedia.

Table 3 Performance comparison with different information source sets (t-test, p-value < 0.000 1)

Information Source Set	Precision	Recall	F_1
$S_{\text{Douban.com Movie}} + S_{\text{Baidu Video}}$	0.257	0.404	0.314
$S_{\text{Douban.com Movie}} + S_{\text{Chinese Wikipedia}}$	0.589	0.706	0.642
$S_{\text{Baidu Video}} + S_{\text{Chinese Wikipedia}}$	0.385	0.591	0.466
$S_{\text{Douban.com Movie}}+S_{\text{Baidu Video}}+S_{\text{Chinese Wikipedia}}$	0.670	0.822	0.738

4.4 Evaluation on evidence combination schemes

To verify the performance of the two-step evidence combination scheme proposed in our method, we compare this scheme with two baselines schemes that integrate the directed and undirected evidence in one step. 1) linear adding, which estimates and combines the weight for each directed evidence similar to the method of Zhu et al. [4] and for each undirected evidence similar to our method; and 2) linear multiplying, which estimates and combines the weight for each directed evidence similar to the method of Zhu et al. [4] and for each undirected evidence similar to our method. Considering the space limitation, we only provide comparison analysis on F_1-score. The two-step evidence combination scheme performs better than the other two baselines (Fig. 3). This result indicates that considering the undirected and directed evidence separately in initial topic hierarchy generation and topic hierarchy improvement is a better way to use both undirected and directed evidence.

4.5 Comparison with state-of-the-art methods

We compare our method against the following state-of-the-art methods by using similar datasets. 1) The method of Heymann [13] is an extensible greedy algorithm for hierarchy generation and uses graph centrality in a similarity graph. The algorithm starts with a root topic similar to our method. The candidate topic terms are added into the topic hierarchy in decreasing order of importance of the term to the topic. The

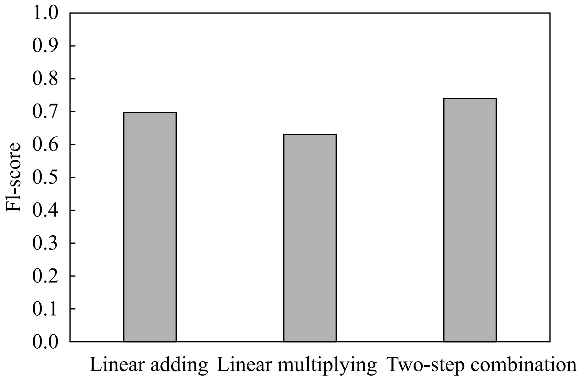

Fig. 3 Performance comparison of different combination schemes for directed and undirected evidence (t-test, p-value < 0.000 1)

probability $p(r(w_i, w_j))$ in our method is employed in Heymann's method for similarity computation. This method decides where to place each candidate term by computing its similarity to every node currently present in the hierarchy, and then adds each candidate as a child of either the most similar node or the root node. 2) The method of Snow [8] uses a probability model to obtain the most probable hierarchy for the given topic term set. We use the gold standard and the probability of a sub-topic relation approximated for our method. 3) The method of Yang [9] designs the information function on the basis of the minimum evolution and abstractness assumptions, and then clusters the candidate topic terms into a topic hierarchy on the basis of the information function. The information function employs all evidence in our method and is trained by using the gold standard. 4) The method of Navigli [11] is a graph-based method that simply uses 0/1 counts for edge weights and is pruned to the hierarchy. We use our heterogeneous evidence to extend this method instead of merely providing is-a relation. 5) Yu [10] extends Yang's method by employing more objective functions including minimum hierarchy discrepancy both from local and global aspects and minimum semantic inconsistency. The information function employs all evidence in our method and is trained by using the gold standard. 6) The method of Zhu [4] is a graph-based method with estimated weights. It incrementally generates topic hierarchy for the given root topic. The root topics and all evidence in our method are used in this method.

Our method significantly outperforms the state-of-the-art methods on all metrics (t-test, p-value < 0.000 1), and the recall is improved between 20.4% and 38.7% in particular (Table 4). We make use of the characteristics of multiple information sources for heterogeneous evidence extraction. And then, we propose a two-step combination scheme to leverage the strengths of undirected and directed evidence. We use undirected evidence to detect the implicit relations for initial topic hierarchy generation and use directed evidence to improve the topic hierarchy. The topic hierarchy improvement is the main reason for the high recall. We give a specific analysis about state-of-the-art methods on F_1-score. Zhu's method comes in second, which indicates that our two-step scheme is a better way to use both undirected and directed evidence. Although taking an incremental greedy algorithm and importance ranking, Heymann's method ranks third because the insertion orders of the candidate topic term is decided by the importance ranking and not by the maximization of the overall relatedness between the candidate topic term and terms currently existing in resultant topic hierarchy. After Heymann's method, Snow's method is also strongly affected by the insertion order of the topic terms. Once an insertion error occurs in one step, such error cannot be corrected in the following steps. Navigli's method simply uses 0/1 values for edge weights rather than the estimated weights to generate a graph and prunes the graph to the hierarchy. This process leads to fewer terms and relations in the resultant topic hierarchy. Therefore, the lowest recall places Navigli's method in the fifth place on F_1-score. Yu's method and Yang's method perform rather poorly partly because they focus more on the structure change than content relevance. Thus, when a topic term appears, it can only be added as a child of any existing term in the hierarchy, whereas the original relation can be broken to create a better structure through heterogeneous evidence on terms of the topic.

Table 4 Performance comparison with state-of-the-art methods (t-test, p-value < 0.000 1)

Method	Precision	Recall	F_1
Heymann's method	0.667	0.527	0.589
Snow's method	0.635	0.504	0.562
Yang's method	0.451	0.445	0.448
Navigli's method	0.626	0.435	0.513
Yu's method	0.538	0.460	0.496
Zhu's method	0.642	0.618	0.630
Our method	0.670	0.822	0.738

4.6 Case study on topic hierarchy construction

In this section, we analyze several examples to demonstrate the performance of the proposed method. For example, {成长 "Growth" → 亲子 "Parenting" → 教育 "Education" → 叫我第一名 "Front of the Class" → 詹姆斯 · 约瑟夫 · 沃尔克 "James Wolk" → 身残志坚 "Broken in Body but Firm in Spirit"} is a branch of topic hierarchy rooted on 成长 "Growth". The directed and undirected evidence from heterogeneous information sources contribute to the result. The directed evidence from Chinese Wikipedia infobox helps detect more types of relations rather than hypernym, such as {叫我第一名 “Front of the Class” → 詹姆斯 · 约瑟夫 · 沃尔克 “James Wolk”}, which stands for the relationship between a film and star. The directed evidence from the Chinese Wikipedia category provides a deeper hierarchy structure, such as {教育 “Education” → 叫我第一名 “Front of the Class”} against the simple {成长 “Growth” → 叫我第一名 “Front of the Class”}. Moreover, the undirected evidence from Douban.com Movie and Baidu Video get more in-depth correlations; 亲子 “Parenting” and 教育 “Education” are more relevant than 成长 “Growth” and 教育 “Education”. Several personalized film reviews, such as 身残志坚 “Broken in Body but Firm in Spirit”, can provide reference to others.

Moreover, our proposed method can discover and organize several movies of similar types, as well as significantly reveal the properties of each movie. For example, in the topic “Growth”, we determine that 阿甘正传 “Forrest Gump” and 舞动人生 “Billy Elliot” are at the level similar to 叫我第一名 “Front of the Class”. They all belong to the branch {成长 “Growth” → 亲子 “Parenting” → 教育 “Education”} of the 成长 “Growth” topic hierarchy. The properties of a film, such as director, actor or actress, and producer, are attached under the film, which are clear and intuitive for users.

4.7 Evaluation on hierarchy update

In this section, we demonstrate how our proposed method can be used to incrementally update the topic hierarchies with newly obtained data. According to the publication date, we divide the data collected from Douban.com Movie, Baidu Video, and Chinese Wikipedia about the top 250 movies of Douban.com Movie into two subsets. We use the data crawled in June 2012 for topic hierarchy construction and those crawled in May 2014 for topic hierarchy update.

By comparing the original topic hierarchies and updated topic hierarchies, our method can effectively detect new emerging topics and organize related terms into the hierarchy rooted on the new topic. As new movies emerge, our method can merge them into the existing topic hierarchies. The updated subset has 43 movies that are different from the original subset. After the update, 5 newly increased topic hierarchies not only contains movies from the list of 43 updated movies, but also include movies from the list of 207 common movies. Among the common 207 movies in the original and updated data sets, approximately 67.2% of the movies belong to more topics than before, whereas the topics of 32.8% movies remain unchanged. For example, 天堂电影院 “Nuovo Cinema Paradiso”, which is a movie of common 207 movies, the assigned topic set is from {成长 “Growth”, 经典 “Classic”} to {成长 “Growth”, 经典 “Classic”, 人性 “Humanity”, 传记 “Biography”, 童年 “Childhood”} after the update process. More topics of a movie can be discovered with the new types of combined movies. This trend indicates that a movie that can be classified into many topics will receive sustained attention. Once we can determine a complete topic picture of a movie, the movie can be recommended to more people corresponding to the topics of interest.

5 Conclusion

This study proposes a topic hierarchy construction method from heterogeneous evidence with real-time update. We learn both topic terms and topic relations entirely from scratch by automatic extraction from multiple Chinese information sources. We make use of the characteristics of multiple information sources for heterogeneous evidence extraction. Furthermore, we present a novel scheme to use the heterogeneous evidence extracted from multiple sources separately in both initial topic hierarchy construction and topic hierarchy improvement by distinguishing them into undirected and directed evidence. Comprehensive experiments demonstrate the effectiveness of our method. Therefore, this study may also help boost the development of traditional hierarchy and ontology construction.

For future studies, we will explore more Chinese folksonomy resources in other domains, such as music, books, sports, and finance. Enhanced performances can be expected when resource exploration and method improvement are combined in the process. Finally, we will apply the generated topic hierarchies to specific tasks, including recommendation and question answering, to advance the development of related tasks.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant Nos. 61273321 and 61133012) and the National 863 Leading Technology Research Project (2015AA015407). Special thanks to Weinan Zhang, Yanyan Zhao, Xiao Ding, Zhenghua Li, Yuhang Guo, Ruiji Fu, Duyu Tang, Muyu Zhang, Luyang Li, Xiaocheng Feng, Hailong Qin, Meishan Zhang, Bo Fu and anonymous reviewers for insightful comments.

References

1. Liu X, Song Y, Liu S, Wang H. Automatic taxonomy construction from keywords. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2012, 1433–

1441
2. Trant J. Studying social tagging and folksonomy: a review and framework. Journal of Digital Information, 2009, 10(1): 1–42
3. Hoffart J, Suchanek F M, Berberich K, Weikum G. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence, 2013, 194: 28–61
4. Zhu X, Ming Z Y, Zhu X, Chua T. Topic hierarchy construction for the organization of multi-source user generated contents. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2013, 233–242
5. Hearst M A. Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th Conference on Computational Linguistic, 1992, 539–545
6. Girju R, Badulescu A, Moldovan D. Learning semantic constraints for the automatic discovery of part-whole relations. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. 2003, 1–8
7. Ming Z Y, Wang K, Chua T S. Prototype hierarchy based clustering for the categorization and navigation of web collections. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2010, 2–9
8. Snow R, Jurafsky D, Ng A Y. Semantic taxonomy induction from heterogenous evidence. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. 2006, 801–808
9. Yang H, Callan J. A metric-based framework for automatic taxonomy induction. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP. 2009, 271–279
10. Yu J, Zha Z J, Wang M, Wang K, Chua T. Domain-assisted product aspect hierarchy generation: towards hierarchical organization of unstructured consumer reviews. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011, 140–150
11. Navigli R, Velardi P, Faralli S. A graph-based algorithm for inducing lexical taxonomies from scratch. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence. 2011, 1872–1877
12. Zhou M, Bao S, Wu X, Yu Y. An unsupervised model for exploring hierarchical semantics from social annotations. In: Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference. 2007, 680–693
13. Heymann P, Garcia-Molina H. Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Technical Report. 2006
14. Angeletou S, Sabou M, Motta E. Semantically enriching folksonomies with FLOR. In: Proceedings of the 1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web. 2008, 1–16
15. Tomuro N, Shepitsen A. Construction of disambiguated folksonomy ontologies using Wikipedia. In: Proceedings of the 2009 Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources. 2009, 42–50
16. Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, (3): 993–1022
17. Tang J, Leung H, Luo Q, Chen D, Gong J. Towards ontology learning from folksonomies. In: Proceedings of the 21st International Joint Conference on Artificial Intelligence. 2009, 2089–2094
18. Bundschus M, Yu S, Tresp V, Rettinger A. Hierarchical Bayesian models for collaborative tagging systems. In: Proceedings of the 9th IEEE International Conference on Data Mining. 2009, 728–733
19. Daud A, Li J Z, Zhou L Z, Zhang L. Modeling ontology of folksonomy with latent semantics of tags. In: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. 2010, 516–523
20. Xue H, Qin B, Liu T. Topical key concept extraction from folksonomy through graph-based ranking. Multimedia Tools and Applications, 2014: 1–19
21. Edmonds J. Optimum branchings. Journal of Research of the National Bureau of Standards B, 1967, 71: 233–240

Han Xue, PhD candidate in computer science, is a student at Harbin Institute of Technology, China. Her interests include information extraction and social computing.

Bing Qin, PhD in computer science, is a professor at the Department of Computer Science of Harbin Institute of Technology, China. Her interests include text mining and natural language processing.

Ting Liu, PhD in computer science, is a professor at the Department of Computer Science and Technology of Harbin Institute of Technology, China. His interests include information retrieval and social computing.

Shen Liu, MS candidate in computer science, is a student at Harbin Institute of Technology, China. His interests include information extraction and natural language processing.