Knowledge-Based Systems 93 (2016) 109–120

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A hybrid term–term relations analysis approach for topic detection

Chen Zhang [a,1,*], Hao Wang [a,b,1], Liangliang Cao [c], Wei Wang [a], Fanjiang Xu [a]

[a] Science and Technology on Integrated Information System Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China
[b] State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China
[c] Yahoo! Labs, 229 West 43rd Street, NY 10036, USA

ARTICLE INFO

Article history:
Received 3 August 2015
Revised 2 November 2015
Accepted 5 November 2015
Available online 14 November 2015

Keywords:
Topic detection
Topic modeling
Semantic relations
Co-occurrence relations
Graph analytical approach

ABSTRACT

Topic detection as a tool to detect topics from online media attracts much attention. Generally, a topic is characterized by a set of informative keywords/terms. Traditional approaches are usually based on various topic models, such as Latent Dirichlet Allocation (LDA). They cluster terms into a topic by mining semantic relations between terms. However, co-occurrence relations across the document are commonly neglected, which leads to the detection of incomplete information. Furthermore, the inability to discover latent co-occurrence relations via the context or other bridge terms prevents the important but rare topics from being detected.

To tackle this issue, we propose a hybrid relations analysis approach to integrate semantic relations and co-occurrence relations for topic detection. Specifically, the approach fuses multiple relations into a term graph and detects topics from the graph using a graph analytical method. It can not only detect topics more effectively by combing mutually complementary relations, but also mine important rare topics by leveraging latent co-occurrence relations. Extensive experiments demonstrate the advantage of our approach over several benchmarks.

1. Introduction

With the rapid development of Web 2.0, the amount of online text is experiencing explosive growth by means of online news media and social media. The rich information within the text data can be utilized to reveal some meaningful trends/topics or the evolution of certain social phenomena, such as the presidential election of the USA. Besides, it can also be exploited for detecting some emergency events or natural disasters, such as *2014 Shanghai Stampede*[2] and *2014 Kangding earthquake*.[3] In general, an event is considered as something non-trivial happening at a specific date/time and in a specific location [1]. A topic can be considered as a kind of "abstract" event which consists of some "concrete" events with semantic relatedness. Being consistent with [1] and [2], "events" and "topics" may be used interchangeably in this paper.

Topic detection is a sub-task of Topic Detection and Tracking (TDT) [2]. It aims at detecting topics or trends from various text corpus such as online media. Topic detection as a fundamental problem of information retrieval can help the decision makers to efficiently detect meaningful topics. Therefore, it has attracted much attention such as public opinion monitoring, decision supporting and emergency management [3–6].

In topic detection research, a topic is usually represented as a set of descriptive and collocated keywords/terms. Initially, document clustering techniques are adopted in topic detection to cluster content-similar documents and extract keywords from clustered document sets as the representation of topics. Currently, most approaches utilize various topic models, a type of generative probabilistic model such as LDA [7], pLSA [8] and their extensions, to detect topics. Among them, LDA has been proved to be a powerful algorithm because of its ability on mining the semantic information from the text data. Terms having semantic relations with each other are collected as a topic.

However, there are two challenges to existing topic model based approaches.

Firstly, as claimed by Sayyadi and Raschid [1], "Current topic modeling methods do not explicitly consider word co-occurrences". "Co-occurrence" means two terms co-occur in the same document. Unfortunately, due to the fact that "Extending topic modeling to include co-occurrence can be a computationally daunting challenge" [1], their proposed graph analytical approach only made an approximation to this extension: they merely took into account co-occurrence information alone while ignoring semantic information. Therefore, how to combine semantic relations and co-occurrence relations to complement each other remains to be a challenge.

* Corresponding author. Tel.: +86 13811513670.
E-mail address: zhangchen@iscas.ac.cn (C. Zhang).

[1] Chen Zhang and Hao Wang have equal contribution to this work.
[2] http://en.wikipedia.org/wiki/2014_Shanghai_stampede.
[3] http://en.wikipedia.org/wiki/2014_Kangding_earthquake.

http://dx.doi.org/10.1016/j.knosys.2015.11.006
0950-7051/

Secondly, existing approaches usually focus on detecting prominent or distinct topics by mining **explicit** semantic relations or **frequent** co-occurrence relations. However, they neglect to uncover **latent** co-occurrence relations. The inability to uncover latent relations prevents the important but rare topics, which are hidden in large scale and noisy data collections, from being detected. Such important rare topics have two attributes: significant for human decision making but rare that cannot be discovered easily [9]. In other words, their features are commonly implicit or latent. Here gives some examples: the latent omens such as the abnormal behaviors of some animals may reveal that the disasters such as earthquake will occur soon; the early incubations of the disease may trigger the subsequent cancer. The reason why such topics are commonly ignored is: they differ from the distinct topics indicating common patterns; besides, they are not outliers or noises in the sense of anomaly detection [9].

How to uncover latent co-occurrence relations? How to discover important rare topics? In Chance Discovery (CD) theory and Idea Discovery (ID) theory, a chance is defined as a rare but important event or situation which has a strong impact on human decision making [9–11]. CD and ID as extensions of Knowledge Discovery have been used to detect chances by uncovering latent co-occurrence relations among terms. In the ID process, text data is analyzed and converted into a term graph by mining co-occurrence relations, where latent co-occurrence relations are uncovered and visualized to capture chances. Latent co-occurrence relations means there may be no frequent co-occurrence relations between two terms (the terms do not frequently co-occur in the same documents); however, **the two terms can be implicitly related/linked by considering the "context" of one of the terms or other bridge terms**. Here "context" denotes neighbors of the term which are strongly interconnected terms in the form of a community (cluster). In other words, **latent co-occurrence relations between two terms cannot be measured in an isolated term-term view; the context of the term should be taken into account**.

To address these challenges, we propose a novel systematic approach to integrate semantic information and co-occurrence information among terms for topic detection. Specifically, the approach fuses multiple types of relations into a uniform term graph by incorporating ID theory with topic modeling method. Firstly, an Idea Discovery algorithm called *IdeaGraph* is adopted to mine co-occurrence relations (especially latent co-occurrence relations) for converting the corpus into a term graph. Then, a semantic relations extraction approach is proposed based on LDA to enrich the graph with semantic information. Lastly, a graph analytical method is presented to exploit the graph for detecting topics.

To the best of our knowledge, the coupling of ID and topic model for topic detection has not been researched until now. As demonstrated in the experiment section, the key superiorities of our approach are as follows:

(1) It can detect topics more effectively to support human decision making by combing mutually complementary relations: semantic relations and co-occurrence relations.
(2) It can mine important rare topics by leveraging latent co-occurrence relations, which may aid human to perceive the topics with great significance.

The rest of this paper is organized as follows. Section 2 introduces related work. Section 3 outlines the framework of our approach. Sections 4, 5 and 6 present the core components of our approach: Section 4 details *IdeaGraph* algorithm for generating the term graph; Section 5 proposes a semantic-relations extraction approach for refining the graph; Section 6 discusses a graph analytical method for extracting the topics from the term graph. Extensive experiments are performed in Section 7 to evaluate the performance of our approach. The conclusion is summarized in Section 8.

2. Related work

2.1. TDT categories

Topic Detection and Tracking (TDT) is an integral part of the DARPA Translingual Information Detection, Extraction, and Summarization (TIDES) program [2]. TDT mainly contains two sub-task: Topic Detection and Topic Tracking. Topic detection (Event detection) aims at detecting novel topics/events from text corpus while topic tracking is dedicated to tracking the evolution of existing topics over temporal dimension. Topic detection has attracted much attention in machine learning, information retrieval and social media modeling [1,3,6,9,12–20]. Specifically, topic detection can be classified into two types: New Event Detection (NED) and Retrospective Event Detection (RED).

NED aims at detecting newly occurred topics/events from online text streams [3]. Rill et al. [21] proposed a system to detect emerging political topics in Twitter. The detected topics can be used to extend existing knowledge bases for better concept-level sentiment analysis. Hou et al. [22] proposed a multifaceted news analysis approach to detect events from online news. They represented news as a link-centric heterogeneous network and formalized news analysis and mining task as link discovery problem. Based on that, they presented a unified probabilistic model for topic extraction and inner relationship discovery within events. Extensive experiments demonstrated the superiority of their approach.

RED is dedicated to discovering the events from the historical corpus in an offline way [23]. Yang et al. [23] proposed an agglomerative clustering algorithm, named augmented Group Average Clustering, to cluster articles into events. They also employed an iterative bucketing and re-clustering model proposed by Cutting et al. [16] to control the tradeoff between cluster quality and computational efficiency. Zeng and Zhang [24] presented a variable space Markov model for topic detection, where several steps based on space computation and a hierarchal clustering algorithm are proposed to tackle the issues of document imbalance and topic transition. We note our paper only addresses RED problem.

2.2. Probabilistic approach

As mentioned previously, some RED approaches employ topic modeling, a type of the probabilistic modeling, to detect topics from the text data. A famous topic model named Latent Dirichlet Allocation (LDA) [7] is a three-level hierarchical Bayesian model. Each document is represented as a finite mixture over an underlying set of latent topics, where each topic is characterized by a distribution over terms. Terms having strong sematic relations with each other are clustered as a topic's features or representation. There are plenty of improved versions of LDA. For example, several knowledge-based topic models have been proposed to incorporate prior domain knowledge from the user to generate coherent topics [25–30]. Among which, Chen and Liu [26] proposed the AMC algorithm, topic modeling with automatically generated Must-links and Cannot-links, to incorporate the knowledge automatically mined from the past learning/modeling results, which can help future learning. Xu et al. [30] used a knowledge based topic model to extract implicit features of product reviews for opinion mining task.

Recently, researchers have shown interest in performing topic detection on social media such as Twitter data using various probabilistic models. Yuan et al. [12] proposed a probabilistic model W^4 (short for What+Who+Where+When) to detect topics with spatial-temporal information from Twitter data. The model can infer the attributes of the topics for a variety of applications, such as user profiling and location prediction. Except the TDT approaches on social media, there also exist some approaches to detect events from online news. Li et al. [13] proposed a probabilistic model for RED. It incorporates both

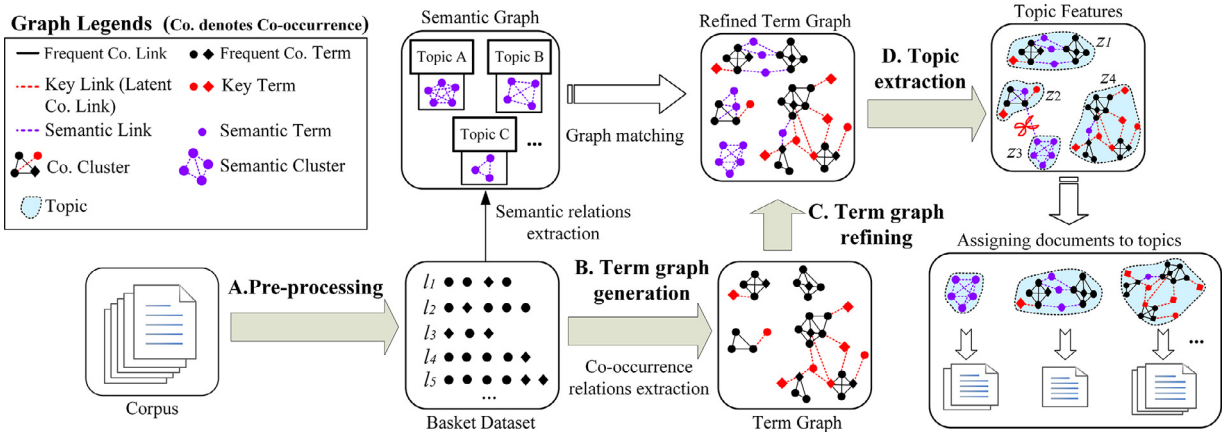

Fig. 1. The framework of the approach.

Table 1
Classification of nodes, links and clusters in the term graph (Co. denotes co-occurrence).

Category	Color	Description
Frequent Co. Link	Black	The link between two terms co-occurring in the same documents frequently.
Frequent Co. Term	Black	The basic elements of a Co. Cluster, interconnected by Frequent Co. Links.
Semantic Link	Purple	The link between two terms sharing coherent semantic information derived by LDA.
Semantic Term	Purple	The basic elements of a Semantic Cluster, interconnected by Semantic Links.
Key Link (Latent Co. Link)	Red	The link between terms *t1* and *t2*, where *t1* and *t2* do not co-occur frequently but can be implicitly linked by considering the context of *t1* or *t2*. The context means all terms in the Co. Cluster where *t1* or *t2* is inside.
Key Term	Red	The hub or bridge term connecting one or more Co. Clusters via Key Links.

content and time information within news text in a unified framework. We note that our approach only refers to RED from online media. It focuses on extracting events (features of keywords) from the text content, while does not address the issue of extracting event information on the temporal, spatial and users dimensions.

As a special variant of topic model, hierarchical topic model [31] can discover the hierarchy of topics which contains various-grained topics. Besides, there are another kind of customized statistical probabilistic models, which concentrate on detecting pre-defined topics such as earthquake and tsunami [4]. However, they are not the pervasive approaches – they fail to detect other kinds of topics.

2.3. Graph analytical approach

Other approaches on topic detection leverage the graph analytical method to detect the topics within the graph or network. Sayyadi and Raschid [1] proposed a graph analytical approach for topic detection. They used a *KeyGraph* algorithm [32] to convert text data into a term graph based on co-occurrence relations between terms. Then they employed a community detection approach to partition the graph. Eventually, each community is regarded as a topic and terms within the community are considered as the topic's features. **Please note that our approach is inspired by their main idea**. However, their approach needs to be improved in two aspects. Firstly, they failed to leverage the semantic information derived from topic model. Secondly, they measured co-occurrence relations in an isolated term-term view: the measurement was limited to term itself. However, the context information of a term was overlooked, which leaded to the inability of measuring latent co-occurrence relations.

Chance Discovery algorithm, *KeyGraph* [32], are dedicated to discover important rare topics called as "chances". In the field of Idea Discovery, our preliminary work – *IdeaGraph* [33–35] can detect chances more effectively by leveraging the ability of human cognition. It is a human-oriented algorithm and can solve a bottleneck of human cognition caused by machine-oriented *KeyGraph* [10]. It converted the text data into a graph using the co-occurrence term-relations. Then the chance can be captured by mining the latent co-occurrence term-relations.

Different from previous research, our approach is a joint framework which combines probabilistic model and graph analytical approach. It can leverage both of them to detect the topics more accurately by integrating semantic information and co-occurrence relations. Besides, its ability on mining latent co-occurrence relations facilitates the discovery of important rare topics.

3. Overview of framework

In this section, we outline the framework of our approach. As shown in Fig. 1, the approach contains four components: (A) Preprocessing; (B) Term graph generation; (C) Term graph refining with semantic information; (D) Topic extraction from the refined term graph.

3.1. Terminology definition of the term graph

A term graph $G = (V, E)$ is generated, where V denotes terms, E denotes term-relations, namely links. Terms are classified into two types based on their occurrence frequencies: frequent term as **dot** which are the top frequent terms up to 30% of the number of terms, and infrequent terms as **diamond** which are the remain 70% terms.

Table 1 shows some terminology definitions. Please note in Fig. 1 Semantic Terms captured by topic model are shown as dots because they are always high frequent. Key Terms does not have strong relations with any single term; but may have strong relations with one or more clusters. So, a Key Link can only be captured by taking context information (cluster) into account. Details of these definitions will be explained in Sections 4 and 5.

Formally, we define a topic as a community/cluster in the term graph, as shown in Fig. 1. In addition, the terms in the cluster are

Table 2
Parameters setting.

Component	Parameter	Description	Value
B	*link_relation*	Threshold of the co-occurrence relation value between terms which are linked in the graph.	0.7
	link_frequency	Threshold of the co-occurrence frequency between terms which are linked in the graph.	7 for Dataset 1 9 for Dataset 2 15 for Dataset 3
	num_keyTerm	The number of Key Terms.	10
	num_keyLink	The number of Key Links.	15
C	*num_semTerm*	The number of Semantic Terms.	30
D	*min_doc2topic_similarity*	Threshold of the similarity between a document *d* and a topic *z* where d can be assigned to *z*.	0.1 for Datasets 1 & 3 0.05 for Dataset 2

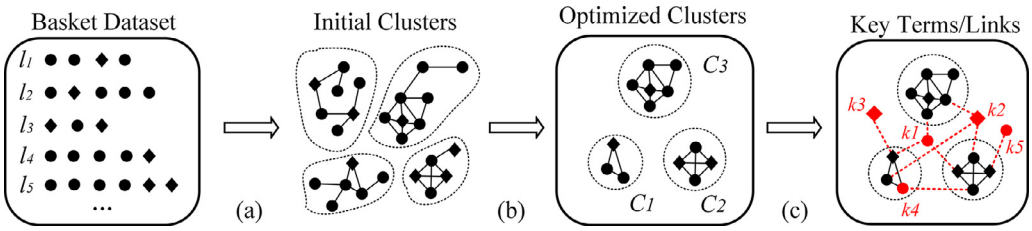

Fig. 2. *IdeGraph* algorithm.

treated as the features, i.e. the keywords, of the topic. Such definition is inspired by the main idea in [1].

3.2. Parameters of the approach

Parameters and their thresholds in the experiment are listed in Table 2. The thresholds are determined by preliminary experiments. The parameters sensitivity will be evaluated in Section 7.5.

3.3. Pre-processing

The main goal of the pre-processing is to filter noise and adjust the data format suitable for the subsequent components. It mainly contains stemming, phase extraction, Part-Of-Speech filtering and stop-word removing. Besides, word segmentation process for Chinese corpus should be performed as the first step of pre-processing. The corpus crawled from the Internet consists of plenty of documents. The documents are pre-processed into a basket dataset *D*. As the input of the subsequent components, *D* is a kind of BOW representation. It consists of some lines representing the sentences of the documents. Each line consists of some terms, i.e. words or phrases. Here gives an example of *D*:

D = *(line 1) term 1, term 2, term 3, term 4.*
(line 2) term 2, term 7, term 5.
(line 3) term 3, term 6, term 9, term 10, term 5.
.....

Next, Sections 4∼6 will describe three core components of our approach.

4. Term graph generation

We adopt an Idea Discovery algorithm named *IdeaGraph*, which was proposed by our previous work [33,35], to convert the basket dataset into a term graph *G* by extracting co-occurrence relations between terms. We take the term graph as the basis to capture topics. As shown in Fig. 2, the algorithm mainly contains three steps: (a) Generating Co-occurrence Clusters; (b) Optimizing the clusters; (c) Extracting Key Terms and key Links.

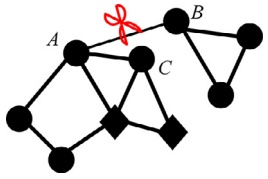

Fig. 3. Cluster pruning.

4.1. Generating Co-occurrence Clusters

Terms having close relations with each other can be linked, eventually forming a **Co-occurrence Cluster**. As shown in Eq. (1), the relation $R(t_i, t_j)$ between two terms t_i and t_j is measured by the maximum of their conjugate conditional probabilities. These two terms are linked by a black line, called as **Frequent Co-occurrence Link**, in the graph *G* if their relation exceeds the threshold *link_relation* and their co-occurrence frequency exceeds another threshold *link_frequency*. In the graph *G*, the terms are called **Frequent Co-occurrence Terms**.

$$\begin{aligned} R(t_i, t_j) &= MAX\left[P(t_i|t_j), P(t_j|t_i)\right] \\ &= MAX\left[\frac{\sum_{l\subseteq L(t_i,t_j)} 1/Ca(l)}{\sum_{l\subseteq L(t_j)} 1/Ca(l)}, \frac{\sum_{l\subseteq L(t_i,t_j)} 1/Ca(l)}{\sum_{l\subseteq L(t_i)} 1/Ca(l)}\right] \end{aligned} \quad (1)$$

where *l* denotes a line of the basket dataset, $Ca(l)$ denotes the capacity of *l*, $L(t_j)$ denotes the set of lines in which t_j occurs, and $L(t_i, t_j)$ denotes the set of lines in which t_i and t_j co-occur.

Please note that the capacity of each line of the basket dataset, i.e. the number of terms in each line, needs to be taken into account. Since if a term appears in a long line, it has more probability to co-occur with other term and vice versa. So we normalize the conditional probabilities in Eq. (1).

4.2. Optimizing the clusters

Clusters should be pruned to enhance the quality, such as removing weak links or partitioning sparse cluster into cohesive subclusters. We prune the clusters by their connectedness: the link *e* is to be pruned when no path connects the two ends of *e* after *e* is pruned. For example, Fig. 3 shows that Link "*A*-*B*" should be pruned while Link "*A*-*C*" should not be pruned.

4.3. Extracting Key Terms and Key Links

The **Key Term** can be extracted by measuring the relations between a term and a Co-occurrence Cluster. The intuition is that a Key Term may have not strong relation with any single term, but it can have strong relations with one or more clusters. For example, a company's executive does not appear very often, but he/she has extensive connections with each head of the departments. Here the executive denotes a Key Term, heads of the departments denote terms in a cluster.

Based on that, the **Key Link** is defined as link between a Key Term and a cluster. Therefore, the cluster as the "context" of the term is utilized to uncover Key Link: latent co-occurrence relations. Based on that, the Key Term as the "bridge" may link various clusters via Key Link.

In this paper, Key Terms are divided into Global Key Term and Local Key Term.

4.3.1. Extracting Global Key Terms

A Global Key Term, such as *k1* or *k2* in Fig. 2, is a term k which has strong relations with all clusters except the cluster where k is inside. Such relation value is called key value and is calculated by Eq. (2).

$$\begin{aligned} Key(k) &= \sum\nolimits_{C_i \in \mathbf{C}, k \notin C_i} R(k, C_i) \\ &= \sum\nolimits_{C_i \in \mathbf{C}, k \notin C_i} \mu R_1(k, C_i) + (1-\mu) R_2(k, C_i) \end{aligned} \tag{2}$$

where C_i denotes any cluster which does not contain the term k, $\mathbf{C}$ denotes the set of clusters, $R(k, C_i)$ denotes the relation between k and C_i, μ ranging from 0 to 1 denotes the weighting factor determining the importance of R_1 and R_2.

In Eq. (2), $R(k, C_i)$ contains two parts which are complementary to each other. The first part is to evaluate the sum of co-occurrence relations between k and each term in C_i, as given by Eq. (3).

$$R_1(k, C_i) = \sum\nolimits_{t \in C_i} R(k, t) \tag{3}$$

The second part is inspired by [36] and can help capturing more Key Terms. It evaluates the degree of variance between two probability distributions: the co-occurrence probability distribution of a term and a cluster, and the unconditional probability distribution of that cluster itself. As explained in [36], such variance degree can be taken as an indicator to reveal the relations between a term and a cluster. It can be calculated by the χ^2– measure in Eq. (4).

$$R_2(k, C_i) = \sum\nolimits_{t \in C_i} [(n_{k,t} - s_k p_t)^2 / s_k p_t] \tag{4}$$

where $n_{k,t}$ denotes the co-occurrence frequency of k and t, s_k denotes the sum of co-occurrence frequencies between t and each term in C_i: $s_k = \sum_{t' \in C_i} n_{k,t'}$, and p_t denotes the unconditional occurrence probability of t in C_i: $p_t = n_t / \sum_{t' \in C_i} n_{t'}$.

All terms are sorted by their $Key(k)$ in a descending order and the *num_keyTerm* terms are labeled as Key Terms and shown as red nodes in the graph G. Key Links are added as red dotted lines to connect the term k and the cluster C_i if their relation is greater than zero. Specifically, we link the term t in C_i to k where among all terms in C_i, t has the largest relations with k.

4.3.2. Extracting Local Key Terms

A Local Key Term, such as *k3, k4* or *k5* in Fig. 2, is a term which has strong relations with one cluster. First, the relations between each term k which does not belong to existing Global Key Terms and each cluster C_i can be measured as $R(k, C_i)$. Then, term-cluster pairs are sorted and top *num_keyLink* pairs are linked as Key Link. Lastly, terms which are linked by newly added Key Link are labeled as Local Key Terms and shown as red nodes in G.

5. Term graph refining with semantic information

We propose a semantic relations extraction approach to enrich the term graph with semantic information. The essence of this approach is how to combine the graph analyzing and the topic modeling for converting point-wise semantic probabilities into the pair-wise semantic relations. The approach mainly contains three procedures: (a) Generating semantic topics; (b) Extracting semantic graph; (c) Merging the term graph and the semantic graph.

5.1. Generating semantic topics

Firstly, we use LDA to build semantic topics. Secondly, we select the feature terms for each topic. Feature terms having relatively high probabilities among the topic's term distribution can well represent the topic's main meaning. We first sort the terms for every topic in a descending order according to the probability distribution of the terms. Then we pick up high probability terms as the feature terms. For each topic, the terms with the probabilities higher than half of maximum of the probability distribution are picked up. Please note this threshold is determined by our preliminary experiment and the experiment indicates our approach is non-sensitive on this parameter.

Please note we do not need to tune topic number K because our approach is insensitive to K. Our strategy is to set a relatively large number as K firstly and then extract the terms having high semantic information from such topics' feature terms. This strategy can filter duplicated topics and meaningless topics automatically.

5.2. Extracting semantic graph

In order to utilize semantic information to enrich the term graph, the semantic topics should be converted into a semantic graph. That is, the high semantic terms should be extracted and the high semantic relations between such terms should be discovered. The procedure consists of three steps:

Firstly, we select candidates for Semantic Terms on each topics. By efficiency issues, we only choose the feature terms generated in the last procedure as the candidates of Semantic Terms, other than the total terms in vocabulary of the basket dataset.

Secondly, we remove the duplicated terms from the candidates. If there are more than one topics having the same term t in their semantic term candidates, only one topic z with the highest term probability $p(t \mid z)$ can remain t as semantic term candidate. Meanwhile, t should be removed from other topic's semantic term candidates. After this step, the semantic term candidates of different topics are exclusive to each other.

Thirdly, we calculate the semantic values for each topic's candidates. Inspired by *tf-idf* formula, we propose a ***tp-izp*** formula to measure the semantic value of any term t in each topic z, which is calculated by Eq. (5):

$$sem(t|z) = p(t|z) \log \frac{p(t|z)}{\sum_{z' \in \mathbf{Z}} p(t|z')} \tag{5}$$

where $\mathbf{Z}$ denotes the set of the semantic topics. In our ***tp-izp*** formula, ***tp*** denotes the term probability, and ***izp*** denotes the inverse topic probability.

To measure semantic relations between terms, we firstly give two hypotheses concluded from our preliminary experiments:

(1) For a topic, the term with higher probability among the term distribution contains more semantic information.
(2) For a topic, the term-pair with closer probability values has closer semantic relations.

Followed by the hypotheses, the semantic relations of any two terms t_i and t_j in the candidates of each topic z is calculated by

Eq. (6):

$$r_{sem}(t_i, t_j|z) = \sqrt{p(t_i|z)p(t_j|z)} \cdot e^{-\frac{|p(t_i|z)-p(t_j|z)|}{\max[p(t_i|z),p(t_j|z)]}} \quad (6)$$

Eq. (6) contains two parts. The first part is proportional to two terms' probability values, which is derived from Hypothesis (1). The second part is inversely proportional to $|p(t_i|z) - p(t_j|z)|$, which is derived from Hypothesis (2). Please note the denominator $\max[p(t_i|z), p(t_j|z)]$ is used for normalization.

All terms in the corresponding candidates are sorted by their $sem(t|z)$ and top *num_semTerm* terms are taken as **Semantic Terms**. Then, the **Semantic Links** between each semantic term-pair within the same topic are measured by $r_{sem}(t_i, t_j|z)$.

Finally, the semantic topics are converted into a semantic graph. In the semantic graph shown in Fig. 2, the purple nodes denote the Semantic Terms and the purple dotted lines denote the Semantic Links. Interconnected Semantic Terms are collectively called a Semantic Cluster.

5.3. Merging the term graph and the semantic graph

As shown in Fig. 1, we merge the term graph with the semantic graph by coupling co-occurrence relations and semantic relations. Finally, new terms are added as Semantic Terms and new links are added as Semantic Links if they did not appear in graph *G*.

6. Topic extraction from the refined term graph

In this section, we present a graph analytical method to exploit the graph for detecting topics. The method mainly contains two procedures: (a) Extracting topic features from the graph; (b) Assigning documents to topics.

6.1. Extracting topic features from the graph

In Fig. 1, we treat each cluster as a topic. A topic denoted as a filled polygon, such as $z1 \sim z4$, can be extracted from the refined term graph. The terms within each polygon are regarded as the features/keywords of the topics.

In some cases, the approach may yield large but sparse clusters having too many terms, which leads to poor topic features. Therefore, as an optional optimized means, we can partition the large cluster into some small but cohesive sub-clusters by using the community detection method mentioned in [1]. For example, topics $z2$ and $z3$ in Fig. 2 do not merge as one topic, although there exists a semantic link connecting them.

6.2. Assigning documents to topics

The procedure contains two steps: (a) Calculating the feature vector for each topic; (b) Measuring the likelihood of each document being assigned to each topic.

6.2.1. Calculating the feature vector for each topic

The keyword weights can be calculated by simply using *tf-idf* formula. However, the graph structure, i.e. the relations between keywords, may be overlooked. Such information are equally important for topic representation. Take Fig. 4 as an example. Fig. 4(a) denotes a topic describing the concept of "data", while Fig. 4(b) denotes a topic discussing the concept of "mining".

Therefore, we calculate the weight of each keyword of the feature vector using Eq. (7):

$$w(t) = \sqrt{|N(t)| \cdot \sum_{t' \in N(t)} R(t, t')} \quad (7)$$

where t denotes a keyword of an topic. $N(t)$ denotes the neighbors of t, in which every term t' is directly linked with t. $R(t, t')$ denotes the relation between t and t'. The equation indicates that the weight of the term t is proportional to two factors: the sum of values of links connecting t, and the number of links connecting t.

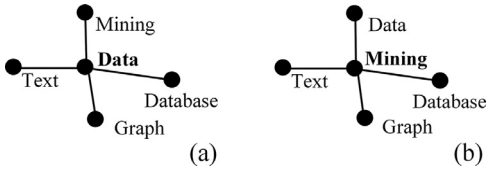

Fig. 4. Different graph structures can represent different topics.

6.2.2. Measuring the likelihood between topics and documents

The likelihood of a document d being assigned to a topic z is measured by Eq. (8).

$$p(z|d) = \sqrt{\frac{\cos(f_d, f_z)}{\sum_{z' \in \mathbf{Z}} \cos(f_d, f_{z'})} \cdot \cos(f_d, f_z)} \quad (8)$$

where $\mathbf{Z}$ denotes the set of topics. f_d and f_z denote the feature vector of d and z. f_d is calculated simply by *tf-idf* formula. Apparently, Eq. (8) reveals the geometrical average of the relative similarity and the absolute similarity. In contrast, the approach proposed by Sayyadi and Raschid [1] only used the relative similarity to measure the likelihood.

Finally, the document d will be assigned to the topic z if and only if two conditions are satisfied simultaneously: (a) $\forall z' \in \mathbf{Z} - z$, $p(z|d) \geq p(z'|d)$; and (b) $p(z|d)$ exceeds the threshold *min_doc2topic_similarity*.

7. Experiment

In this section, we conduct the experiments and report the results to evaluate the effectiveness and efficiency of our approach on capturing topics by comparing them with the results achieved from several benchmarks.

7.1. Datasets

We employ three corpora as the experiment datasets. Dataset 1 containing the human-annotated ground truth is for evaluating the performance and the robustness of our approach. Dataset 2 is also for evaluating the performance, where the evaluation is more subjective because there is no objective ground truth for Dataset 2. Dataset 3 is for testing the running efficiency of our approach.

Dataset 1 is collected from Sohu News [37], which contains 1600 Chinese news documents. These documents cover 8 categories, such as Vehicle, Finance and Society. Every domain contains 200 documents. The total time span for 8 categories is from Sep 1st of 2011 to Nov 25th of 2014. In each category, the time range is approximately five months. The vocabulary contains 94,981 unique terms after preprocessing.

Dataset 1 has 22 human-identified topics/events as the ground truth, which is annotated by 4 annotators. The reliability of agreement between annotators is tested to ensure the validity of the annotations. Annotators independently reviewed the titles and the abstracts to determine the number of topics, the keywords of each topic, and the corresponding set of documents belonging to each topic. Then they voted the score on each topic. As a result, the ground truth is obtained. The topic with the maximum granularity has almost 180 documents, while the minimum has only five.

Dataset 2 is crawled from Twitter, which consists of 74,323 tweets. These tweets belong to 700 users. That is, approximately 106 tweets per user. The time span is from Sep 1st to Sep 30th of 2014. The

vocabulary contains 381,610 unique terms after stop-word removing and POS filtering. Unlike Dataset 1, Dataset 2 have not human-annotated ground truth.

Dataset 3 is a "bigger" version of Dataset 1. It is collected from 12 Chinese news websites and contains almost 150,000 documents covering 10 categories [37]. The time span is from Jan 1st of 2012 to Nov 30th of 2014. The vocabulary contains about 958,000 unique terms after pre-processing.

7.2. Evaluated approaches

We evaluate two variants of our approach as follows.

(1) ***IG***. A variant of our approach that only employs *IdeaGraph* mentioned in Section 4 to generate the term graph, but without the graph refining component mentioned in Section 5.
(2) ***LDA-IG***. Our approach equipped with *IdeaGraph* and LDA.

We compare them with the following benchmarks.

(1) ***KG***. *KeyGraph*, a graph analytical approach proposed by Sayyadi and Raschid [1].
(2) **LDA**. Latent Dirichlet Allocation, a classic unsupervised topic model proposed by Blei et al. [7].
(3) **AMC.** A knowledge-based topic model proposed by Chen and Liu [26].
(4) ***kNN***. The k-Nearest Neighbor clustering algorithm proposed by Allan et al. [3].
(5) **GAC**. The augmented Group Average Clustering algorithm proposed by Yang et al. [23].

Please note here we employ LDA as the topic model of our approach because it is easy to implement. We will deploy various extended topic models such as AMC for better performance in the future. For Dataset 1, AMC uses 15,000 documents randomly sampled from Dataset 3 as the train corpus to mine prior knowledge.

All these approaches are equally equipped with pre-processing mentioned previously. The parameters of them are set in a comparable standard. That is, we performed the approaches using various parameters and we chose the best results to be compared. For example, the parameters of LDA/AMC are set to$\alpha = 0.02$; $\beta = 0.5$; $K = 22$ for Dataset 1 and 20 for Dataset 2. Please note 22 is the true #topics of Dataset 1. We train LDA/AMC using 2000 iterations with an initial burn-in of 200 iterations. Please note *KG, IG*, and *LDA-IG* share some parameters, which control the construction process of the co-occurrence graph. These parameters are set equally in each approach. The parameters of *LDA-IG* will be discussed in Section 7.5.

7.3. Evaluation on the performance of topic detection

We employ Dataset 1 and Dataset 2 to evaluate the performance of the approaches.

Dataset 1 has the ground truth – a set of reference topics and a corresponding set of documents per reference topic. So the approaches are evaluated directly by **Precision, Recall** and **F1 score**:

Precision on features extracting of each extracted topic z is calculated as $num(K_i)/num(K)$, where K denotes the set of keywords of z, and K_i denotes the intersection between K and K_0. K_0 denotes the set of keywords of the human-annotated topic z_0 which has the most number of keywords in common with z from the ground truth. **Recall** on features extracting of z is calculated as $num(K_i)/num(K_0)$. **Precision** on document assigning of z is calculated as $num(D_i)/num(D)$, where D denotes the set of approach-calculated documents being assigned to z, and K_i denotes the intersection between D and D_1. D_1 denotes the set of documents being assigned to the human-annotated topic z_1 which has the most number of documents in common with z from the ground truth. **Recall** on document assigning of z is calculated as $num(D_i)/num(D_1)$. **F1 score** is calculated as the harmonic mean of corresponding precision and recall.

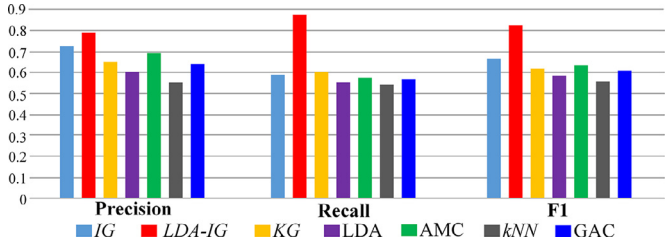

Fig. 5. Performance of the approaches on topic features extracting using Dataset 1 (Values are calculated using the average across all detected topics, similarly hereinafter in Table 3, and Fig. 9).

Table 3
Frequency of the topic keywords extracted by the approach using Dataset 1.

Approaches	IG	LDA-IG	KG	LDA	AMC	kNN	GAC
Frequency	60	147	134	258	249	198	225

Besides, **Frequency** – the average frequency of topics' keywords, can also be used to reveal the ability on discovering important rare topics. The lower frequency indicates the more latent topics are captured.

However, Dataset 2 has not human-annotated ground truth, so it is impossible to directly evaluate precision and recall. Here we employ **Pseudo-Precision** and **Pseudo-Inverse-Recall** proposed by Sayyadi and Raschid [1] as the evaluation metrics.

Therefore, Dataset 1 is used for evaluating topic features extracting and document assigning[4]; Dataset 2 is used for only evaluating document assigning, similar as [1] mentioned.

7.3.1. Performance on topic features extracting

Results on Dataset 1 are summarized in Fig. 5 and Table 3. Fig. 5 indicates the precision, recall and F1 score of *LDA-IG* on topic features extracting exceeds that of other approaches. Table 3 shows the topic features captured by *LDA-IG* are of relatively lower-frequency than that of LDA, AMC, *kNN* and GAC. Therefore, *LDA-IG* can extract important topics more effectively, even if they do not appear frequently.

We give the detailed demonstrations on the improvements of *LDA-IG* in F1 score:

Firstly, the improvement of *LDA-IG* is about 24% over *IG*. As mentioned previously, it is expected that *LDA-IG* can leverage semantic information to detect more topics than *IG*.

Secondly, the improvement of *LDA-IG* is about 33% over *KG*, 41% over LDA, 28% over AMC, 48% over *kNN* and 37% over GAC. Although the precision of AMC is better than LDA because AMC leveraged the mined prior knowledge to produce more coherent topics than LDA, such prior knowledge did not contain the co-occurrence information. Therefore, the improvement of AMC on Recall is slight compared with LDA.

The superiority of *LDA-IG* is mainly benefited from taking hybrid term relations into account when extracting topics. However, the other benchmarks used either co-occurrence relations or semantic relations, which led to the detection of incomplete information. We take *KG* as a demonstration in Fig. 6 to explain such superiority:

(1) Key/Semantic Terms can enrich the features of the topic. These terms can ease cognitive load and support human to better perceive the importance of the topic. For example, topic $z1$ in Fig. 6(a) contains 6 keywords while $z1'$ in Fig. 6(b) contains 3 keywords. In general, $z1$ is more informative and easy-understood than $z1'$.

[4] As mentioned in Section 6, the task of event detection mainly contains two subtasks: Extracting event features in Section 6.1, and assigning documents to events in Section 6.2.

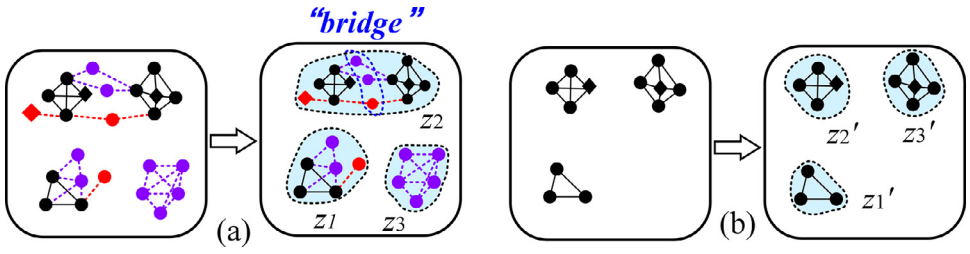

Fig. 6. Comparison between *LDA-IG* and *KG* on extracting topics. (a) denotes *LDA-IG*; (b) denotes *KG*.

Table 4
List of the part of topics detected by *LDA-IG*, *KG* and LDA using Dataset 1.

Approaches	Topics	Categories	English translations of the representative keywords (keywords with low feature weights are omitted)
LDA-IG	$z1$	Vehicle	Car, 4S store, Engine, Power, X-Car.net, Turbine, Vehicle, Brand, Design, Machine filter, Insurance costs
	$z2$	Finance	Fund, Bond, Bank, Currency translation, Participation, Economy, Market, Investment, Rate increasing/ cutting
	$z3$	Disaster	Earthquake, China Earthquake Administration, Bearing wall, "*Kangding*", "*Daofu*", Mudslide, Gallows
	$z4$	Society	Beijing, Shanghai, Livelihood, Rising, Real estate, Limited Purchasing Bill, Housing price
	$z5$	Politics	Relationship, International, Nation, Cooperation, Politics, News/Media, Government, Taiwan, Japan, Mainland
	$z6$	Health	Hospital, Eat, Food, Patient, Treatment, Disease, Doctor
KG	$z1'$	Vehicle	Luxury Car, Guide price, Quoted price, Change, Remark, 4S store, Models, Current car
	$z2'$	Vehicle	Infiniti, Vehicle.com.cn, Tabulation, X-Car.net
	$z3'$	Vehicle	Turbine, Direct Injection, Engine, Peak, Power, Torque, Cow
	$z4'$	Finance	Fund Bank, Management, Rights, Services, Currency translation, Greentown, Kowloon Price
	$z5'$	Disaster	Bridge, Convenience, Reconstruction, House, Bearing wall, Assistance
	$z6'$	Society	Beijing, Shanghai, Rising, Area, Guangzhou, Livelihood
LDA	$z1''$	Vehicle	Vehicle, Brand, Design, Models, Engine, Products, Users, Consumers, Internet, Cell phone
	$z2''$	Finance	Market, Fund, Rate increasing, Rate cutting, Bank, Money, Investment, Economy, Company, Currency, Business
	$z3''$	Finance	Development, Enterprise, Company, Problem, Market, Service, Country, Job, Beijing, Product, Project
	$z4''$	Travel	Tourism, Tourist, Scenic spot, Travel agency, Hotel, Goods, Golf, Swimming, Tour guide, Travel
	$z5''$	Health	Hospital, Patient, Doctor, Eat, Food, Treatment, Disease, operation, Eyes, Medicine
	$z6''$	Politics	Politics, News/Media, Government, Nation, Relations, Military, Economic, International, Cooperation, Rival

(2) Key/Semantic Terms can help merging sub-clusters having latent co-occurrence relations or semantic relations with each other via Key/Semantic Links. For example, incomplete topic $z2'$ and $z3'$ in Fig. 6(b), which may belong to the same theme, are merged into complete $z2$ in Fig. 6(a). In this case, the three Key/Semantic Terms circled by the blue dotted line are regarded as "bridge" terms, which connect two sub-clusters.
(3) Semantic Terms can help to detect more topics. For example, topic $z3$ in Fig. 6(a) are detected by using semantic information; while it is overlooked in Fig. 6(b).

Table 4, Figs. 7 and 8 give the detailed examples to demonstrate the superiority mentioned above. These examples show a part of topics extracted by *LDA-IG*, *KG* and LDA, respectively. We exclude AMC, *kNN* and GAC from comparison to satisfy space limitations because their results are comparable with LDA. As for *LDA-IG* in Table 4, the black terms, the red terms and the purple terms denotes the *Frequent Co-occurrence Terms*, the *Key Terms* and the *Semantic Terms* in the term graph, respectively.

Here shows some **CASE STUDIES** according to the examples:

(1) Topic $z1$ in *LDA-IG* belongs to category Vehicle. The Key Terms and Semantic Terms can help merging the three sub-topics, which denote as $z1'$, $z2'$ and $z3'$ in *KG*.
(2) Topic $z2$ in *LDA-IG* belongs to category Finance. The Key terms and Semantic terms can enrich the features of the topic, which leads to a more easily understandable topic than $e4'$ in *KG*. Such terms can facilitate human to quickly perceive the significance of the topics.
(3) Topic $z3$ in *LDA-IG* belongs to category Disaster and is about an earthquake occurred in *Kangding* and *Daofu* of China, November, 22, 2014. Similarly, $z4$ belongs to category Society which refers to the quickly-rising housing prices in Beijing and Shanghai during 2014.
Such topics as important rare topics are cognitively essential for human comprehension, interpretation, and decision making. For example, $z3$ indicates the specific locations of the earthquake and can help human to perceive the potential secondary disasters such as mudslide. Topic $z4$ may motivate human to infer that the government will issue *Limited Purchasing Bill* to suppress the rising tendency of housing prices.
However, they cannot be captured by LDA due to their infrequency (the number of news articles separately reporting $z3$ and $z4$ is five and eleven). In addition, $z5'$ and $z6'$ in *KG*, which are similar with $z3$ and $z4$, are hard to comprehend because of poor feature terms. Hence, they cannot well support human decision making.
(4) Topics $z5$ and $z6$ belong to categories Politics and Health, respectively. They are captured by using semantic relations between terms. Especially when comparing with $z6''$ in LDA, the key terms in $z5$ enrich the feature of $z5$ and make it more informative. That is, the content of $z5$ is focused on Taiwan issue while the content of $z6''$ is more general. However, they cannot be captured by *KG* due to the neglect of the semantic information.

7.3.2. Performance on document assigning

(1) Dataset 1

As shown in Fig. 9, *LDA_IG* outperforms other approaches on document assigning. The advantage can be analyzed in two aspects:

(1) *LDA_IG* uses Eq. (8) to measure the likelihood of each document being assigned to each topic. There are two factors in Eq. (8): the relative similarity and the absolute similarity. The geometrical average of two factors in Eq. (8) can improve the precision

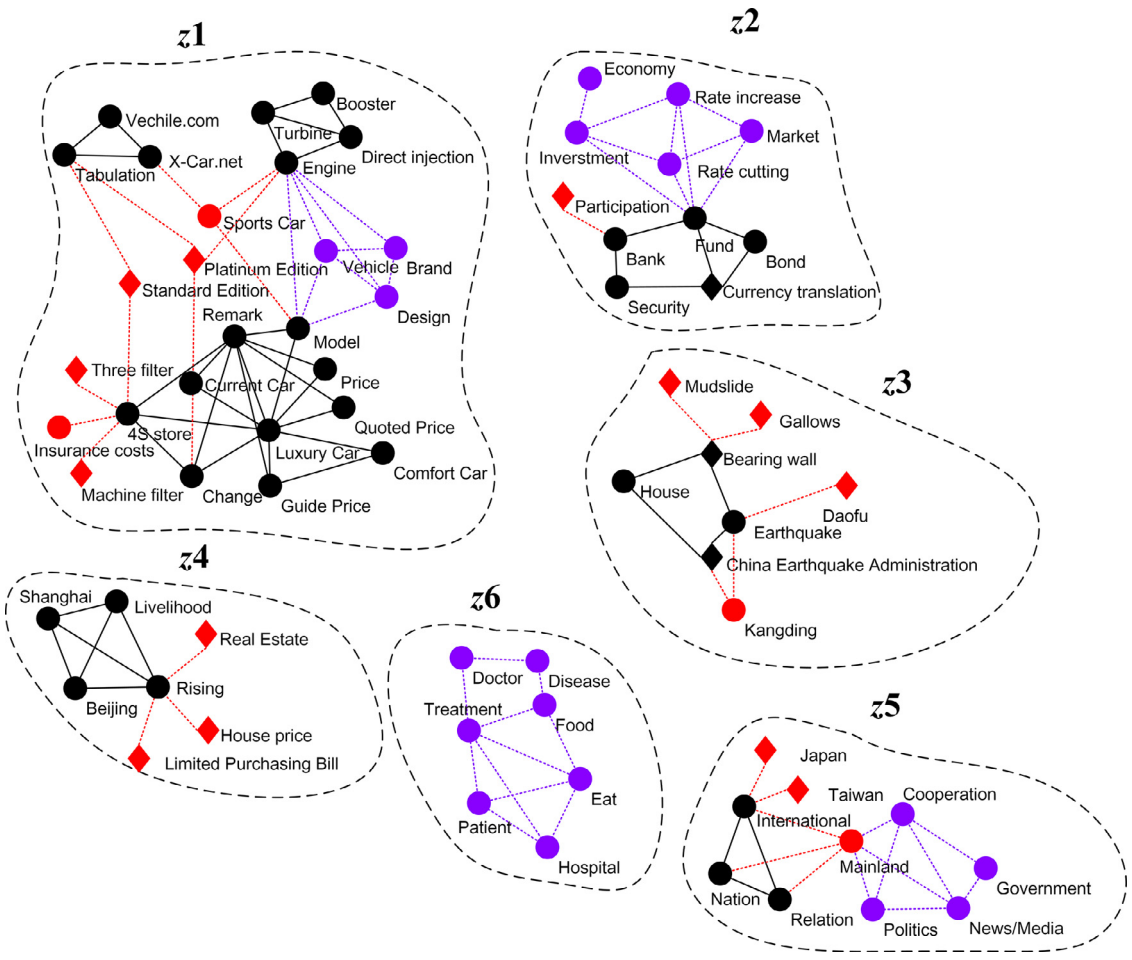

Fig. 7. Sub-graph of the term graph generated by *LDA-IG* using Dataset 1.

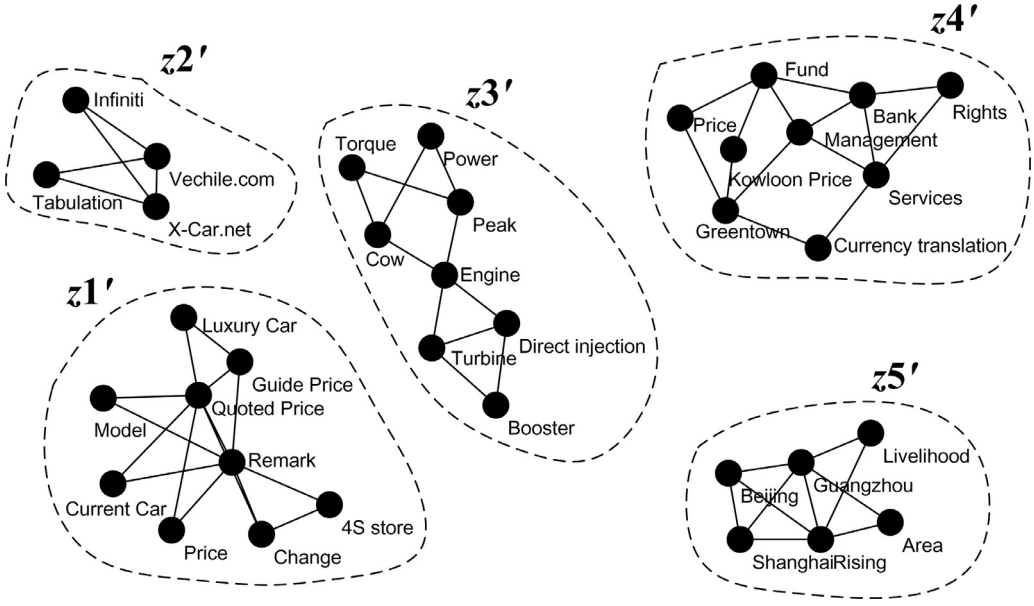

Fig. 8. Sub-graph of the term graph generated by *KG* using Dataset 1.

Table 5
Pseudo-Precision of the approaches on document assigning using Dataset 2.

Approaches	#Docs	#Evaluated docs	%Not relevant	%Somewhat	% Relevant	%Highly relevant
LDA	6441	2000	53.9	30.0	9.4	6.7
KG	3992	1382	17.8	42.0	14.9	25.3
LDA-IG	5942	1661	**26.5**	**24.1**	**20.8**	**28.6**

Table 6
Pseudo-Inverse-Recall of the approaches on document assigning using Dataset 2.

Approaches	#Evaluated docs	%Not relevant	%Somewhat	% Relevant	%Highly relevant
LDA	1200	63.3	32.0	3.6	1.1
KG	1200	55.0	24.6	10.0	10.4
LDA-IG	1200	**59.5**	**22.9**	**6.5**	**11.1**

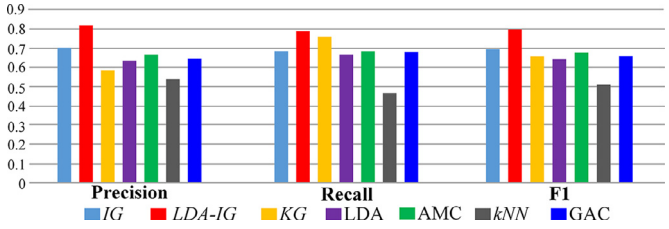

Fig. 9. Performance of the approaches on document assigning using Dataset 1.

of assigning documents to corresponding events. In contrast, *KG* only employed the relative similarity to measure the likelihood.

(2) As mentioned in Section 7.3.1, the feature terms of the detected topics from *LDA_IG* are of relatively higher-quality and more reasonable than that of other approaches. Better features can also improve the performance of document assigning.

(2) Dataset 2

We compare the performance of *LDA_IG*, *KG* and LDA on document assigning using Dataset 2. AMC, *kNN* and GAC are excluded from comparison to satisfy space limitations because their results are comparable with LDA. Besides, *KG* and LDA are representative that they use term relations to detect topics.

We strictly refer to the evaluation method mentioned by Sayyadi and Raschid in Section 7.3.2 of [1] and follow the evaluation process of them as much as possible. Specifically, **Pseudo-Precision** and **Pseudo-Inverse-Recall** are employed as the evaluation metrics[5]. The details of these metrics are omitted owing to space limitations. The best 20 topics are evaluated for the approaches; for each topic a random subset of documents with the number between 50 and 150 are chosen.

The evaluation results are revealed in Tables 5 and 6. The results reveal that *LDA-IG* gets the highest pseudo-precision and the lowest pseudo-inverse-recall, which are consistent with the results on Dataset 1. Please note Dataset 2 consists of Twitter corpus, which are full of noisy and diverse short text. So the performance of the approaches on Dataset 2, especially LDA, is worse than that on Dataset 1.

7.4. Evaluation on the running time

Dataset 3 is employed to evaluate the approach efficiency. Fig. 10 shows the running times of the approaches under various numbers of processed documents. The number varies from 1000 to almost 150,000. The approaches are performing on a PC with an Intel Xeon @3.0 GHz processor and a 16 GB memory running Windows 8.

Please note the efficiency of AMC and GAC are not evaluated because that AMC contains a training process for obtaining prior knowledge from a large-scale corpus, which is rather time-consuming. Besides, the time complexity of GAC varies significantly under various parameter values.

We stop running *KG* after 75,000 documents, since it is very slow. It appears that the running time of *KG* increases in a nonlinear trend. The reason is that the strategy of *KG* on graph construction is to build large clusters first and then partition them into small clusters; while the time complexity of community detection algorithm employed by *KG* for partition is $O(n^3)$, where n denotes the number of unique terms. In contrast, the strategy of *LDA-IG* is to build small clusters first and then merging them into cohesive ones, which is relatively efficient.

On the other hand, the running time of remain approaches increases in a linear trend. In Fig. 10(b), the running efficiency of *LDA-IG* is comparable as LDA and *kNN*, which indicates *IdeaGraph* algorithm is very efficient on graph construction. Apparently, the advantage of *IdeaGraph* algorithm over LDA can be explained that *IdeaGraph* algorithm does not depend on a learning process.

7.5. Evaluation on the parameters sensitivity

As shown in Fig. 11, the parameters sensitivity of *LDA-IG* is evaluated. We run *LDA-IG* on Dataset 1 for different parameter values and report the performance on the evaluated parameters. The descriptions of the parameters are listed in Table 2.

Considering the roles of different parameters, Fig. 11(c) shows the performance of document assigning while the rest figures show the performance of topic features extracting. When evaluating each parameter, the other parameters are fixed to their best values in Table 2. Please note the hyper-parameters in topic model module are fixed as $\alpha = 0.02$; $\beta = 0.5$, which are equal to LDA's hyper-parameters.

The first two parameters *link_relation* and *link_frequency*: The performance slightly drops for relatively small values for *link_relation* and *link_frequency*. However, the performance drops significantly for very large values of *link_relation* and *link_frequency*. This is expected that these parameters control the performance of graph constructing. So these parameters should not be set too big.

The parameter *min_doc2topic_similarity*: Fig. 11 shows that *LDA-IG* is considerably robust to the value of *min_doc2topic_similarity*.

The last three parameters *num_keyTerm, num_keyLink* and *num_semTerm*: Fig. 11 shows that *LDA-IG* is also robust to the value of these three parameters.

Eventually, our approach has been proved to be rather robust on its parameters setting.

[5] Please refer to Section 7.3.2 of [1].

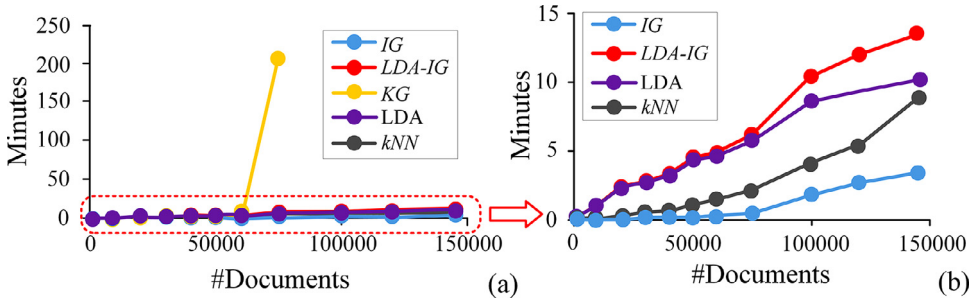

Fig. 10. Running time of the approaches. As denoted by red dotted box in (a), (b) is the subset of (a), which removes the results of *KG* (Values are calculated using the average across five running).

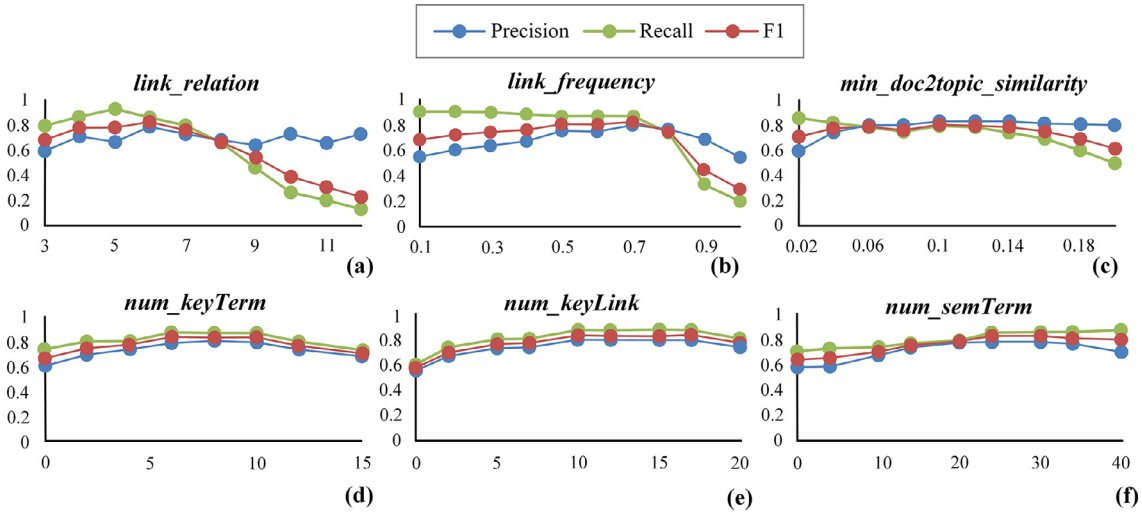

Fig. 11. Sensitivity analysis for parameters of *LDA-IG*.

8. Conclusion

This paper proposes a hybrid relations analysis approach integrating semantic relations and co-occurrence relations for topic detection. The approach seamlessly incorporates topic modeling with chance discovering to capture semantic relations and co-occurrence relations. Such coupled relations facilitate detecting topics more effectively. In addition, the approach uncovers latent co-occurrence relations to detect important rare topics by taking the context information into account. It provides us a comprehensive perspective to perceive the significance of topics. To the best of our knowledge, the incorporation of such relations for topic detection has not been researched until now. The extensive experiments demonstrate the superiority of our approach by comparing with several benchmark approaches.

Further work can focus on the extension of the approach by using the incremental clustering algorithm for NED task to fit the needs in the era of Big Data. Besides, how to track the evolution of topics over time using the hybrid relations analysis approach also remains to be a challenge. Lastly, we will extend the approach to represent the topic hierarchy, where events as the basic elements of topics need to be detected.

Acknowledgment

This work is supported by National Basic Research Program of China (2013CB329305), Natural Science Foundation of China (61303164, 61402447, 61502466) and Beijing Natural Science Foundation (9144037). This work is also sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

References

[1] H. Sayyadi, L. Raschid, A graph analytical approach for topic detection, ACM Trans. Internet Technol 13 (2) (2013) 4:1–4:23.
[2] J. Allan, J.G. Carbonell, G. Doddington, J. Yamron, Y. Yang, Topic detection and tracking pilot study, in: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998, pp. 194–218.
[3] J. Allan, R. Papka, V. Lvrenko, On-line new event detection and tracking, SIGIR (1998) 37–45.
[4] T. Sakaki, M. Okazaki, Y. Matsuo, Tweet analysis for real-time event detection and earthquake reporting system development, IEEE Trans. Knowl. Data Eng. 25 (4) (2013) 919–931.
[5] Z. Yang, Y. Liu, D. Hou, T. Feng, Y. Wei, J. Zhang, P. Huang, G. Zhang, Water quality event detection based on Multivariate empirical mode decomposition, SMC (2014) 2663–2668.
[6] T. Sabbah, A. Selamat, M.H. Selamat, R. Ibrahim, and H. Fujita, Hybridized term-weighting method for dark web classification. Neurocomputing (2015) doi:10.1016/j.neucom.2015.09.063.
[7] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, Adv. J. Mach. Learn. Res. 3 (2003) 993–1022.
[8] T. Hofmann, Probabilistic latent semantic analysis, UAI (1999) 289–296.
[9] Y. Ohsawa, Chance discoveries for making decisions in complex real world, New Gen. Comput. 20 (2) (2002) 143–163.
[10] H. Wang, Y. Ohsawa, Idea discovery: a scenario-based systematic approach for decision making in market innovation, Expert Syst. Appl. 40 (2) (2013) 429–438.
[11] H. Wang, Y. Ohsawa, Y. Nishihara, Innovation support system for creative product design based on chance discovery, Expert Syst. Appl. 39 (5) (2012) 4890–4897.
[12] Q. Yuan, G. Cong, Z.Y. Ma, A.X. Sun, N.M. Thalmann, Who, Where, When and What: discover spatio-temporal topics for Twitter users, KDD (2013) 605–613.
[13] Z. Li, B. Wang, M. Li, W.Y. Ma, A probabilistic model for retrospective news event detection, SIRIR (2005) 106–113.
[14] F. Chen, D.B. Neill, Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs, KDD (2014) 1166–1175.
[15] L. Hong, A. Ahmed, S. Gurumurthy, A.J. Smola, K. Tsioutsiouliklis, Discovering geographical topics in the Twitter stream, WWW (2012) 769–778.

[16] D.R. Cutting, D.R. Karger, J.O. Pedersen, J.W. Tukey, Scatter/gather: A cluster-based approach to browsing large document collections, SIGIR (1992) 318–329.
[17] J.A. Hartigan, M.A. Wong, Algorithm AS 136: a k-means clustering algorithm, Appl. Stat. 28 (1) (1979) 100–108.
[18] C. Li, A. Sun, A. Datta, Twevent: segment-based event detection from tweets, CIKM (2012) 155–164.
[19] Z. Zhao, S. Feng, Q. Wang, et al., Topic oriented community detection through social objects and link analysis in social networks, Knowl. Based Syst. 26 (2012) 164–173.
[20] K. Ravi, V. Ravi, A survey on opinion mining and sentiment analysis: tasks, approaches and applications, Knowl. Based Syst. 89 (2015) 14–46.
[21] S Rill, D Reinel, J. Scheidt, et al., Politwi: early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis, Knowl. Based Syst. 69 (2014) 24–33.
[22] L. Hou, J. Li, Z. Wang, et al., NewsMiner: Multifaceted news analysis for event search, Knowl. Based Syst. 76 (2015) 17–29.
[23] Y. Yang, T. Pierce, J.G. Carbonell, A study on retrospective and on-line event detection, SIGIR (1998) 28–36.
[24] J. Zeng, S. Zhang, Variable space hidden Markov model for topic detection and analysis, Knowl. Based Syst. 20 (7) (2007) 607–613.
[25] D. Andrzejewski, X. Zhu, M. Craven, Incorporating domain knowledge into topic modeling via Dirichlet forest priors, ICML (2009) 25–32.
[26] Z. Chen, B. Liu, Mining topics in documents: standing on the shoulders of big data, KDD (2014) 1116–1125.
[27] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, R. Ghosh, Exploiting domain knowledge in aspect extraction, EMNLP (2013) 1655–1667.
[28] A. Mukherjee, B. Liu, Aspect extraction through semi-supervised modeling, ACL (2012) 339–348.
[29] X. Fu, K. Yang, J.Z. Huang, L. Cui, Dynamic non-parametric joint sentiment topic mixture model, Knowl. Based Syst. 82 (2015) 102–114.
[30] H. Xu, F. Zhang, W. Wang, Implicit feature identification in Chinese reviews using explicit topic mining model, Knowl. Based Syst. 76 (2015) 166–175.
[31] D.M. Blei, T.L. Griffiths, M.I. Jordan, J.B. Tenenbaum, Hierarchical topic models and the nested Chinese restaurant process, Advances Neural Information Processing Systems 16, MIT Press, Cambridge, MA, 2004.
[32] Y. Ohsawa, N.E. Benson, Y. Masahiko, KeyGraph: automatic indexing by cooccurrence graph based on building construction metaphor, ADL (1998) 12–18.
[33] H. Wang, F.J. Xu, Y. Ohsawa, IdeaGraph: a graph-based algorithm of mining latent information for human cognition, SMC (2013) 952–957.
[34] C. Zhang, H. Wang, F. Xu, X. Hu, IdeaGraph Plus: a topic-based algorithm for perceiving unnoticed events, ICDMW (2013) 735–741.
[35] C. Zhang, H. Wang, W. Wang, F.J. Xu, An improved IdeaGraph algorithm for discovering important rare events, SMC (2014) 3290–3295.
[36] M. Yutaka, I. Mitsuru, Keyword extraction from a single document using word co-occurrence statistical information, AAAI (2003) 392–396.
[37] SohuNews, 2014. Retrieved December 30, 2014: http://news.sohu.com/ (accessed 18.07.15).