Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

A semi-explicit short text retrieval method combining Wikipedia features

Pu Li [a,*], Tianci Li [a], Suzhi Zhang [a,*], Yuhua Li [a], Yong Tang [b,*], Yuncheng Jiang [b,*]

[a] Software Engineering College, Zhengzhou University of Light Industry, Zhengzhou 450000, China
[b] School of Computer Science, South China Normal University, Guangzhou 510631, China

ARTICLE INFO

Keywords:
Semi-explicit semantic
Feature selection
Short text retrieval
Wikipedia

ABSTRACT

With the advantages such as openness, interactivity, immediacy, and simplicity, the large number of short text data appear in the Web information space. Considering the short length, little information, sparse features and irregular grammar, the traditional information analyzing and retrieval technologies cannot deal with short text effectively. In view of the above problems, in this paper a new short text retrieval method based on the current mainstream semantic knowledge source, Wikipedia, is proposed. To be specific, a semantic feature selection algorithm is proposed to return the top k most relevant Wikipedia concepts as the whole vector space for a given short text. Thus, by analyzing the topic information of the semantic features contained in Wikipedia concepts, we propose some formulas to determine the association coefficient list between different components of the corresponding positions in two different feature vectors. On this basis, a new semantic relatedness assessment method under this lower dimensional semantic space is designed. According to computing and sorting the semantic relatedness between user queries and the target short text, a novel semi-explicit short text retrieval method combining Wikipedia concept feature and the corresponding topic information is proposed. Lastly, based on the experimental results on twitter subsets, we verify that our proposal has advantages over other some current retrieval methods on *MAP*, *P@k* and *R-Prec*, and can return more valid results.

1. Introduction

With the development of the Internet, short message, Twitter, microblogs and WeChat have gained rapid popularity and become the main platforms for people to publish information and conduct social activities. The popularity of these new types of social media have led to the emergence of a large amount of short text data (such as BBS comments, instant message records, news headlines, product reviews, etc.) in the Web information space. Considering the short text information published in above social media is usually the emergency, real-time news and so on, which can attract the most attention, short text has become an important source for users to obtain the information they need (Gan and Wang, 2015).

Unlike traditional long text, the content organization of short text usually does not follow the grammatical structure, which contains many abbreviations and slangs. These characteristics make the short text has sparse feature information and limited semantic representation. Obviously, it is difficult for the machine to obtain enough information content under a limited context to understand and analyze the short text (Alsmadi and Gan, 2019; Wang et al., 2016). This is also why the traditional information retrieval technologies cannot process short text effectively (Kalloubi et al., 2016). In view of the above problems, this paper takes the semantic relatedness as the starting point and uses Wikipedia as the external semantic knowledge source to study the short text retrieval technology. According to analyzing the implicit topic information between explicit feature concepts in Wikipedia, a semantic feature selection and semantic relatedness computation method is proposed. On this basis, a semi-explicit short text retrieval method combining Wikipedia features is proposed. Finally, the feasibility and effectiveness of the method are verified by experimental tests.

The remainder of this paper is organized as follows. Section 2 briefly reviews the state of the art in short text understanding and retrieval technologies. Section 3 provides our research questions and research methodology. In Section 4 we design a semantic feature selection algorithm which returns the top k most relevant Wikipedia concepts as the feature vector for the target short text. Then, we obtain short text semantic relatedness by computing the topic information of each feature concept. By combining the above semi-explicit Wikipedia features, a new short text retrieval model is proposed in Section 5. Section 6 details the experiments that evaluate the effectiveness of our method and reports the analysis of results. Finally, we draw our conclusion and outline the future work in Section 7.

2. Related work

Early short text retrieval method was mainly originated from the traditional long text retrieval technology. The key idea that adopted

[*] Corresponding authors.
E-mail addresses: superlipu@163.com (P. Li), zhsuzhi@zzuli.edu.cn (S. Zhang), ytang4@qq.com (Y. Tang), ycjiang@scnu.edu.cn (Y. Jiang).

https://doi.org/10.1016/j.engappai.2020.103809
Received 23 October 2019; Received in revised form 1 May 2020; Accepted 9 July 2020
Available online 16 July 2020

was keyword matching strategy based on the "Bag-of-words" model, such as tf–idf, BM25 and probability model (Chen et al., 2010). However, the retrieval methods based on the "Bag-of-words" model often ignore the problems of polysemy, synonym and variant, and also overlook the extension of the implicit semantics of the short text (Azad and Deepak, 2019). So, the problem of lacking semantic feature information in short text cannot be solved well, which leads to the unsatisfactory retrieval results. In order to solve these problems, many researches begin to use external knowledge sources to extend the information of the short text, so as to better understand the deep semantic information contained in the short text (Alsmadi and Gan, 2019; Huang et al., 2017; Li et al., 2017a; Nasir et al., 2019; Qu et al., 2018).

The current short text understanding methods are mainly divided into three semantic models: implicit semantic model, semi-explicit semantic model and explicit semantic model (Wang et al., 2016). The implicit semantic model maps short texts into an implicit vector in semantic space. The meaning of each dimension of the vector cannot be interpreted intuitively by humans but can only be used for machine processing. The representative works of implicit semantic model are latent semantic analysis (LSA) (Deerwester et al., 1990), hyperspace analogue to language mode (HAL) (Lund and Burgess, 1996) as well as the neural language model (NLM) (Bengio et al., 2003; Mikolov et al., 2010) and paragraph vector (PV) (Le and Mikolov, 2014) which are developed on this basis. The semi-explicit semantic model also adopts vector space to represent short text. However, unlike implicit semantic model, each dimension of the vector under the semi-explicit semantic model is a topic which is usually a group of words or concepts. Therefore, semi-explicit semantic model is also called topic model (Li et al., 2018b; Zhang and Zhong, 2016). The earlier topic model was based on LSA and was called probabilistic LSA (PLSA) (Hofmann, 2004). This line of thinking leads Blei et al. to proposed a more complete model, latent Dirichlet allocation (LDA) (Blei et al., 2003), which solved the problem that PLSA lacks of the priori distribution of hypothetical topics. On this base, Cuong et al. make a further research about the problem of overfitting of probabilistic topic models on short noisy text (Cuong et al., 2019). Although the general meaning of the corresponding dimension can be inferred from the target topic, these inferred semantics are still uncertain. Different from the above two models, explicit semantic model focuses on transforming short text into a vector space that both humans and machines can understand. Under this model, each dimension of a short text vector has clear semantics, usually is a specific concept. In this way, people can easily understand the vector and make further adjustments and optimizations. There are two common frameworks for this model: explicit semantic analysis (ESA) (Gabrilovich and Markovitch, 2007; Li et al., 2017b; Mohamed and Oussalah, 2019) and conceptualization (Song et al., 2011; Wang et al., 2015).

Based on the different short text semantic understanding models shown above, some new short text retrieval methods have been proposed. In the simplest case, some researchers matched the term in the short text with the concept in the knowledge source (Liu et al., 2010; Mendes et al., 2010). While Tang et al. pointed out that this matching strategy cannot solve the problem of ambiguity concept (Tang et al., 2012). To overcome these limitations, Meij et al. linked n-gram terms in text to Wikipedia with various features (n-gram features, concept features and Twitter features) (Meij et al., 2012). The supervised learning algorithm used in this method provides much more semantics for the short text and improves the accuracy of matching. Unlike the aforementioned term mapping based strategies, other short text retrieval methods used topic feature based strategies (Abel et al., 2011; Lau et al., 2012) or latent semantic based strategies (Ke et al., 2012; Vicient and Moreno, 2015) extracted from context. Based on the nearest neighbor cluster fusion, Liang et al. used Wikipedia to realize short texts representation and retrieval (Liang et al., 2014). While references Lu et al. (2014) and Ensan and Al-Obeidat (2019) implemented Ad-Hoc's short text query and entity selection tasks by introducing Wikipedia concepts and Tie-breaking strategies respectively. In addition, some researchers had also proposed some microblog retrieval methods based on time distribution information and topic features (Han et al., 2016; Xiao et al., 2017).

In the last two years, some new advances had been made in the research and application about short text. With the combination of multiple enhancement graphs and LDA models, Liu et al. proposed a retrieval method for social short texts in 2018 (Liu et al., 2018). Chu et al. applied a topic diffusion strategy and proposed a new short text clustering method (Chu et al., 2017). Li et al. proposed a new noise filtering method for short text preprocessing, which improved the correctness of short text modeling and retrieval (Li et al., 2018a). In 2019, Chen et al. conducted a research and experimental analyses about short text topic discovery and retrieval strategies based on LDA and non-negative matrix factorization (NMF) (Chen et al., 2019). While other researchers using LDA to construct the understanding model for short text (Chen and Ren, 2017; Tajbakhsh and Bagherzadeh, 2019; Zhu et al., 2019). They pointed out that LDA model can effectively construct the topic space for short text and improve the effect of short text retrieval. Moreover, some other studies focus on the short text similarity measuring (Bekkali and Lachkar, 2019; Huang et al., 2019; Yao et al., 2018) as well as clustering (Kozlowski and Rybinski, 2019; Qiang et al., 2019; Song et al., 2019; Yang et al., 2019).

However, reference Kalloubi et al. (2016) indicated that the existing methods usually assume that all concepts are independent and equivalent with each other without considering the subordination relationship between concepts. Therefore, this horizontal semantic extension can still not reflect the implicit information contained in the short text. Starting from this position, this paper investigates the topic information of feature concepts extracted from Wikipedia and construct a more reasonable semantic model for short texts. By using this new model, the semantic features of short text can be extended vertically, and the retrieval effect can be improved effectively.

3. The research questions and methodology

In this section, we will analyze the limitations of the existing short text retrieval methods and present our research questions and the corresponding methodology.

3.1. Motivation

As pointed out in Kalloubi et al. (2016), in the process of semantic extension for short text, the existing short text understanding and retrieval methods usually assume that all concepts are independent and equivalent with each other when constructing the vector space. For example, if the term "Intelligence" appears in short text, the Wikipedia concept "Artificial Intelligence" in Wikipedia can be discovered as the semantic extended feature information by applying word segmentation and mapping techniques. While some other concepts which have semantic relevance but not syntax similarity such as "Cybernetics" or "Machine Learning" may be omitted. These two terms are considered to be useless as "Food" and "King Kong" but this is not the case. This will lead to the lack of partial semantic information in the semantic extension process for short text understanding, thus affecting the final performance of retrieval.

3.2. The main objective and contributions

To counter the above problem, in this paper, we will propose a novel semi-explicit short text retrieval method combining Wikipedia concept features and the corresponding topic information. The key variables and main contributions of our research are as follows:

(1) We further analyze the limitations of the current vector model-based information processing methods. In order to construct a low-dimensional vector space with more relevant semantics, we define a

notion as the formally representation about the top k most relevant Wikipedia concepts and design a feature selection algorithm to return the top k most relevant Wikipedia concepts.

(2) To assess the semantic relatedness between two short texts under a relatively low-dimensional vector space without any extension, we present some formulas to compute the association coefficient list and transform different feature vectors into the same semantic space by analyzing the corresponding topic information of the feature concepts in Wikipedia.

(3) Using the association coefficient list, we provide new method to assess the semantic relatedness between the user query and the target short text and design a framework as the formalized expression of our semi-explicit short text retrieval model.

(4) On this basis, we use the benchmark as well as some wildly used metrics to evaluate the effectiveness of our new retrieval method by comparing with some of the most representative similarity methods.

4. Wikipedia-based semantic feature selection and relatedness computation

It is easy to see from the aforementioned research status that the common steps adopted by the existing short text retrieval methods are as follows: Firstly, the short text is modeled by different mapping methods (such as implicit, semi-explicit or explicit semantic model). Then according to the characteristics of different vector spaces, different strategies are used to compute the degree of association so as to obtain the retrieval result. Along this line, this paper first proposes a semantic relatedness computation method for short text based on topic characteristics of Wikipedia concept.

4.1. Short text semantic feature selection based on Wikipedia concept

Firstly, we will analyze the reason why we chose Wikipedia as external knowledge source to realize short text retrieval as well as the experiment about tweets in this paper.

(1) From the domain of knowledge: as explained at the beginning of this paper (see Sections 1 and 2), considering the short length, little information, sparse features and irregular grammar of short text, many existing short text-oriented researches generally introduce external knowledge sources to extend the semantic of short text so as to realize short text intelligent understanding. Although Wikipedia has different genre from short text (e.g. tweets), as we all know, Wikipedia, a free encyclopedia in all languages, is easily the largest, domain independent, most widely used, and fastest growing encyclopedia in existence. These characteristics are exactly in line with tweets. Considering that tweets are short texts published by different users for different issues, they also contain information in various domains, so we need to find a knowledge source that covers as much as possible. In this way, we can effectively deal with all kinds of tweets and reduce the influence of knowledge in specific domain on short text understanding.

(2) From the structure of knowledge: there are also some other open-source knowledge sources, such as DBpedia (Bizer et al., 2009; Lehmann et al., 2015), YAGO (Hoffart et al., 2013; Mahdisoltani et al., 2014; Suchanek et al., 2008) and so on. But these knowledge sources mainly contain structured data composed of RDF triples, and most of them are extracted from Wikipedia. In other words, they are a subset of Wikipedia, respectively. Although this kind of structured data has a clear logical structure, it filters out a lot of literal content when building the graph structure, and only retains the core concepts and corresponding relationships. While Wikipedia is a kind of semi-structured knowledge source, its data forms are more diverse. Wikipedia not only contains structured data such as infobox and Wikipedia Category Graph (WCG), but also text data such as article and gloss. This kind of diverse data representation is more conducive to the modeling of text objects. Especially for short text, it contains sparse feature information and limited semantic representation, so if we still use structured knowledge source, there will be a lot of semantic information which cannot be matched and computed, which leads to the problem of semantic sparsity cannot be solved well. However, Wikipedia can extend the semantics of short text through its text information effectively.

(3) In addition, during the experimental evaluation stage, by introducing data preprocessing and noise filtering technologies, the core information both in Wikipedia and short text will be retained, which can better ensure the accuracy.

To sum up, even though the content styles are different, by comparing the knowledge characteristics between Wikipedia and short text (e.g. tweets), we still choose Wikipedia as external knowledge source. Besides, the experimental results given in this paper also show that short text can be effectively understood and retrieved by analyzing its semantic feature information in Wikipedia.

Along this line, by using Wikipedia as external knowledge source, we firstly improve the existing ESA algorithm and obtain the top k most relevant Wikipedia concepts as basic semantic features to construct explicit semantic models for short texts. On this basis, a relatively low-dimensional vector space is constructed.

For clarity of presentation, we give a formal definition about the information contained in Wikipedia pages.

Definition 1 (*Wikipedia Article, Wikipedia Concept*)**.** Given a Wikipedia page, the *Wikipedia article* denoted as *Art*, is the abstract shown before the content list. The corresponding *Wikipedia concept*, denoted as *Con*, is the title of the article as well as the Wikipedia page. The formal representation of Wikipedia concept *Con* is defined as the following five tuple:

Con $= \langle Redirection, Wikipedia\ glosses, Anchors, Categories, InfoBox\rangle$, where

(1) *Redirection* is the redirection or disambiguation links, which can be seen as a set of synonyms e.g. $\{Con_1, Con_2,..., Con_n\}$;
(2) *Wikipedia glosses* is the first paragraph of Wikipedia article;
(3) *Anchors* is the set of *Con* annotated as anchor texts in Wikipedia article (i.e., labels of internal hyperlinks);
(4) *Categories* indicates the category of *Con*, which are displayed at the bottom of the corresponding page in the form of links e.g. $\{Cat_1, Cat_2,..., Cat_m\}$;
(5) *InfoBox* displays the main information of feature concepts in tabular form.

The formal representation of Wikipedia concept "*Car*" obtaining from Wikipedia in December 2018 is shown in Fig. 1.

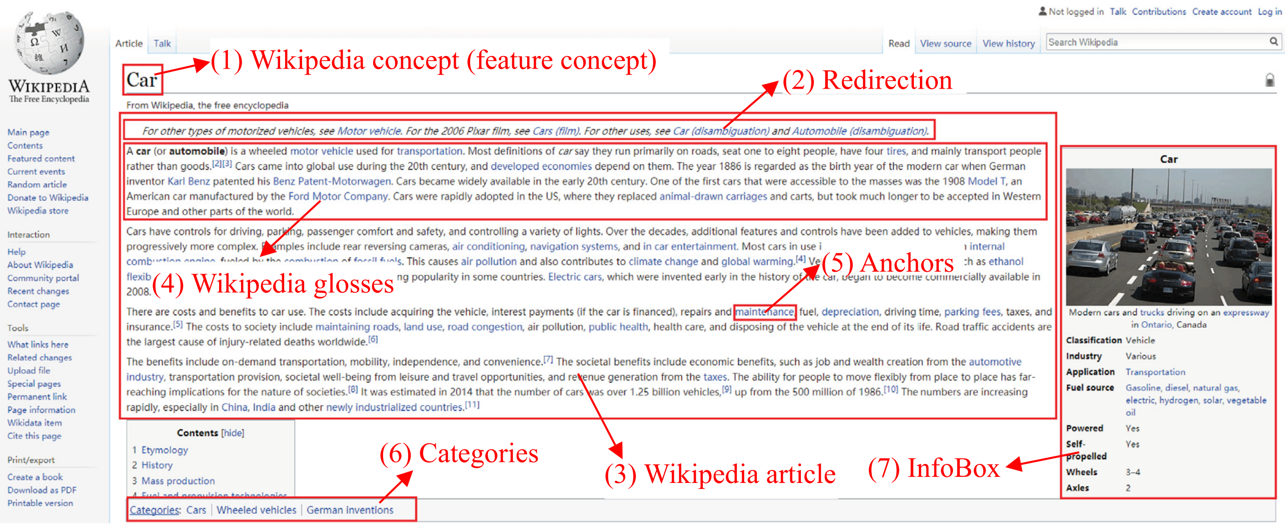
As we all know, ESA uses latter association-based method to assign semantic interpretation to terms and text fragments. With an inverted index, ESA maps each term or text into a set of Wikipedia concepts $\{Con_1, Con_2,..., Con_n\}$. Every Wikipedia concept Con_i is represented as an attribute dimension of the vector about the given target term or text *t*. The weight of each dimension w_i represents the strength of association between *t* and Con_i using tf–idf scheme (Salton and McGill, 1986; Yahav et al., 2018) (e.g. $t = \{w_1, w_2,..., w_n\}$, where *n* denotes the number of concepts in Wikipedia). The specific formula of tf–idf scheme can be seen at formula (1).

$$w^{(Art)}(s_i) = tf\left(s_i, Art\right) \cdot \log \frac{n}{df(s_i)} \tag{1}$$

where $w^{(Art)}(s_i)$ is the tf–idf weight of stem s_i towards article *Art*, *tf*$(s_i$, *Art*) is the number of occurrence of stem s_i in *Art*, *df*(s_i) is the number of articles containing stem s_i, *n* is the total number of all articles.

After that, ESA uses Cosine metric to compare semantic relatedness of $\langle t_1, t_2\rangle$. Fig. 2 illustrates the process of ESA which is used from the paper of Gabrilovich and Markovitch (Gabrilovich and Markovitch, 2007).

From the computing process shown in Fig. 2, we can easily conclude that the inverted index should contain all concepts embodied in Wikipedia and the high-dimensional vector space will be established.



Car

From Wikipedia, the free encyclopedia

For other types of motorized vehicles, see Motor vehicle. For the 2006 Pixar film, see Cars (film). For other uses, see Car (disambiguation) and Automobile (disambiguation).

A **car** (or **automobile**) is a wheeled motor vehicle used for transportation. Most definitions of *car* say they run primarily on roads, seat one to eight people, have four tires, and mainly transport people rather than goods.[2][3] Cars came into global use during the 20th century, and developed economies depend on them. The year 1886 is regarded as the birth year of the modern car when German inventor Karl Benz patented his Benz Patent-Motorwagen. Cars became widely available in the early 20th century. One of the first cars that were accessible to the masses was the 1908 Model T, an American car manufactured by the Ford Motor Company. Cars were rapidly adopted in the US, where they replaced animal-drawn carriages and carts, but took much longer to be accepted in Western Europe and other parts of the world.

Cars have controls for driving, parking, passenger comfort and safety, and controlling a variety of lights. Over the decades, additional features and controls have been added to vehicles, making them progressively more complex. Examples include rear reversing cameras, air conditioning, navigation systems, and in car entertainment. Most cars in use i ı internal combustion engine, fueled by the combustion of fossil fuels. This causes air pollution and also contributes to climate change and global warming.[4] Ve ch as ethanol flexib ng popularity in some countries. Electric cars, which were invented early in the history of the car, began to become commercially available in 2008.

There are costs and benefits to car use. The costs include acquiring the vehicle, interest payments (if the car is financed), repairs and maintenance, fuel, depreciation, driving time, parking fees, taxes, and insurance.[5] The costs to society include maintaining roads, land use, road congestion, air pollution, public health, health care, and disposing of the vehicle at the end of its life. Road traffic accidents are the largest cause of injury-related deaths worldwide.[6]

The benefits include on-demand transportation, mobility, independence, and convenience.[7] The societal benefits include economic benefits, such as job and wealth creation from the automotive industry, transportation provision, societal well-being from leisure and travel opportunities, and revenue generation from the taxes. The ability for people to move flexibly from place to place has far-reaching implications for the nature of societies.[8] It was estimated in 2014 that the number of cars was over 1.25 billion vehicles,[9] up from the 500 million of 1986.[10] The numbers are increasing rapidly, especially in China, India and other newly industrialized countries.[11]

Contents [hide]
1 Etymology
2 History
3 Mass production
4 Fuel and propulsion technologies

Categories: Cars | Wheeled vehicles | German inventions

Car

Modern cars and trucks driving on an expressway in Ontario, Canada

Classification	Vehicle
Industry	Various
Application	Transportation
Fuel source	Gasoline, diesel, natural gas, electric, hydrogen, solar, vegetable oil
Powered	Yes
Self-propelled	Yes
Wheels	3–4
Axles	2

Fig. 1. Wikipedia page for "*Car*" (https://en.wikipedia.org/wiki/Car).

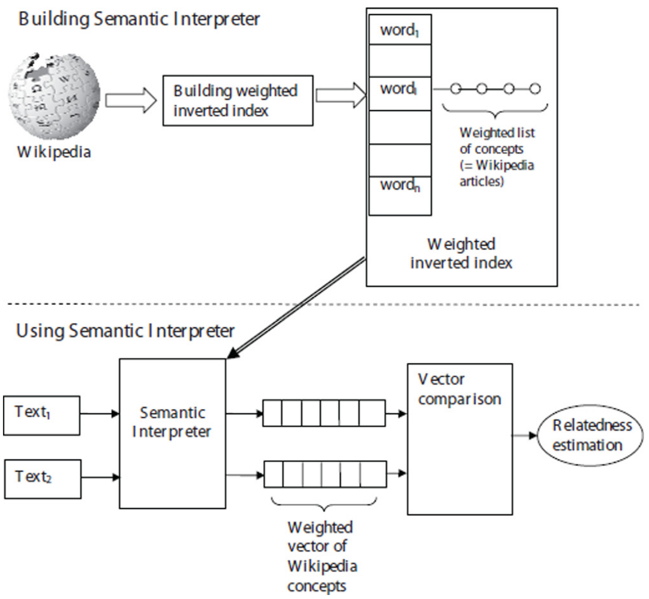

Building Semantic Interpreter
Wikipedia
Building weighted inverted index
word1
wordi
wordn
Weighted list of concepts (= Wikipedia articles)
Weighted inverted index
Using Semantic Interpreter
Text1
Text2
Semantic Interpreter
Vector comparison
Relatedness estimation
Weighted vector of Wikipedia concepts

Fig. 2. The process of ESA algorithm.

In this vein, due to the continuous increase of the size of Wikipedia, the dimensions of concept vector for a given term or text are extremely high. According to the latest version of Wikipedia in 2019, it contains more than 5.7 million concepts, so for each given term or text fragment, the dimension of the feature vector generated by ESA naturally exceeds 5.7 million.

However, without loss of generality, the vector space for a given term or text may become a sparse matrix. Because, in most cases, a stem just appears in quite a few articles with a nonzero weight. So the computation of Cosine over two high-dimensional sparse vectors will meet many 0 values and do a lot of unnecessary cost. Even worse, some very low weights appearing at some dimensions may cause unexpected side-effect on the semantic relatedness assessment even though the values are not null. Obviously, in this high-dimensional sparse matrix, the efficiency and effectiveness will be greatly affected when using ESA to compute the semantic relatedness and then to understand and retrieve short texts.

To cope with the issues above, we will introduce a sorting strategy for the inverted index generated by ESA model. For a given term or text fragment, the Wikipedia concept Con_i in concept vector $\{Con_1, Con_2, \ldots, Con_n\}$ is sorted by its tf–idf weight. Only the concepts with higher rank (e.g. the top 1000) should be identified as "relevant". The higher mark one weight scores, the more relevant the concept is. Those lower-ranked concepts whose weights are too low or 0 value may be worthless even counterproductive.

Along this line, we define a notion as the formally representation for the above description.

Definition 2 (*Relevant Concept List*)**.** Let d be a short text, The *related concept list* corresponding to d, denoted as $RL_{Top\text{-}k}$, is defined as $L = \langle C_1, \ldots, C_k \rangle$. Each element C_i in L is a two tuple (i.e. $C_i = \langle Con_i, w_i \rangle$), where Con_i is a Wikipedia concept defined in Definition 1, w_i is the corresponding tf–idf weight of d for Con_i, and $k \in \{1, 2, 3 \ldots\}$. For any two elements in L, $C_i = \langle Con_i, w_i \rangle$ and $C_j = \langle Con_j, w_j \rangle$, satisfy the following conditions:

(1) if $i \neq j$, then $Con_i \neq Con_j$;
(2) if $i < j$, then $w_i \geq w_j$.

The key steps of the algorithm for obtaining $RL_{Top\text{-}k}$ are as follows:

In what follows, let us see a simple example to make a further explanation for Definition 2 as well as Algorithm 1.

Example 1. Let us consider the pair of terms (also can be seen as two short texts) given in a well known benchmark WordSim-353, available online,[1] $\langle car, automobile \rangle$. Here we set the threshold $k = 10$ for the length of $RL_{Top\text{-}k}$. According to Definition 2, we can obtain the top 10 most relevant concepts for "*car*" and "*automobile*" respectively by using Algorithm 1. The result is shown in Table 1.

Obviously, from Table 1 we can conclude the following results:

(1) $RL_{(car)Top\text{-}10} = \langle\langle \text{Light car}, 0.392749548 \rangle, \langle \text{European Car of the Year}, 0.3878404498 \rangle, \langle \text{R179 (New York City Subway car)}, 0.3852950335 \rangle, \langle \text{List of fastest production cars}, 0.3816963732 \rangle, \langle \text{Concept car}, 0.3816413879 \rangle, \langle \text{Car dealership}, 0.3677532077 \rangle, \langle \text{European Touring Car Cup}, 0.3647055626 \rangle, \langle \text{Australian Production Car Championship}, 0.3614672124 \rangle, \langle \text{Appendix J Touring Cars}, 0.3599415421 \rangle, \langle \text{1941 Indianapolis 500}, 0.3477960527 \rangle\rangle$;

(2) $RL_{(automobile)Top\text{-}10} = \langle\langle \text{Automobile Magazine}, 0.5446857214 \rangle, \langle \text{List of automobile sales by model}, 0.4905870557 \rangle, \langle \text{Automobile repair shop}, 0.4037753046 \rangle, \langle \text{Car dealership}, 0.3984113038 \rangle, \langle \text{Center}$

[1] http://www.cs.technion.ac.il/gabr/resources/data/wordsim353/wordsim353.html.

```
Algorithm 1: Extracting $RL_{Top\text{-}k}$ from Wikipedia for the target short text $d$
Input:
    $d$: a short text
    $sl$: a stop word list
    $k$: a threshold
    $CV=\{Con_1, Con_2, \ldots, Con_n\}$: the Wikipedia concept vector
Output:
    $RL_{Top\text{-}k}$ : the relevant concept list for $d$
Begin:
    [Data preprocessing and noise filtering]
    $d$=Eliminate($d$, $sl$)    // Eliminate the meaningless words in $d$ by $sl$
    $d$=Parser($d$)    // Detect the key phrases in $d$ using Shallow Parser
    $d$=Stem($d$)    //Make normalization for effective terms in $d$ using Porter Stemming
    Foreach $Con_i \in CV$
        $Con_i$=Eliminate($Con_i$, $sl$)    // Eliminate the meaningless words in $Con_i$ by $sl$
        $Con_i$=Parser($Con_i$)    // Detect the key phrases in $Con_i$ using Shallow Parser
        $Con_i$=Stem($Con_i$)    //Make normalization for effective terms in $Con_i$ using Porter Stemming
    End
    Foreach $Con_i \in CV$
        $w_i$=TfIdf($d$, $Con_i$)    // Assign the weight $w_i$ of $d$ for $Con_i$ using tf-idf
        $Index$=Add($Index$, $<Con_i, w_i>$)    // Construct the inverted index between $d$ and $CV$
    End
    Foreach $<Con_i, w_i> \in Index$
        $RL$=Sort($Index$, $w_i$)    // Sort $<Con_i, w_i>$ in $Index$ according to $w_i$ for $d$
    End
    $RL_{Top\text{-}k}$=Select($RL$, $k$)    // Select the top $k$ most relevant concepts according to the threshold $k$
Return $RL_{Top\text{-}k}$
```

Table 1
RL_{Top-k} (k = 10) for "*car*" and "*automobile*".

ID	RL_{Top-k} (k = 10) for "*car*"			RL_{Top-k} (k = 10) for "*automobile*"		
	Wikipedia ID	Con_i	w_i	Wikipedia ID	Con_i	w_i
1	7258157	Light car	0.392749548	1183732	Automobile Magazine	0.5446857214
2	459478	European Car of the Year	0.3878404498	10324705	List of automobile sales by model	0.4905870557
3	16921217	R179 (New York City Subway car)	0.3852950335	5339220	Automobile repair shop	0.4037753046
4	4833382	List of fastest production cars	0.3816963732	**2562877**	**Car dealership**	**0.3984113038**
5	390885	Concept car	0.3816413879	6237341	Center console (automobile)	0.3982055187
6	**2562877**	**Car dealership**	**0.3677532077**	6723676	West Riding Automobile Company	0.3959100842
7	22860903	European Touring Car Cup	0.3647055626	9698053	List of U.S. Routes in New York	0.3824058473
8	12923145	Australian Production Car Championship	0.3614672124	24814175	Automobile Manufacturers Association	0.3761255443
9	19058194	Appendix J Touring Cars	0.3599415421	415724	Antique car	0.3718522489
10	5196759	1941 Indianapolis 500	0.3477960527	14268944	National Automobile Dealers Association	0.3695437908

console (automobile), 0.3982055187⟩, ⟨West Riding
Automobile Company, 0.3959100842⟩, ⟨List of U.S. Routes in New
York, 0.3824058473⟩, ⟨Automobile Manufacturers Association,
0.3761255443⟩, ⟨Antique car, 0.3718522489⟩, ⟨National Automobile
Dealers Association, 0.3695437908⟩⟩.

4.2. Short text semantic relatedness assessment based on topic features

It is well known that computing the distance between two vectors in a vector space model usually uses Cosine metric. Therefore, many vector model-based information processing methods usually use "dot product" to compute the semantic relatedness, such as ESA, SVM, etc. A very important precondition for using Cosine metric is that two vectors must have the same dimensions for their lengths as well as the attribute of each dimension.

However, since the sorting strategy is introduced when acquiring $RL_{Top\text{-}k}$, for two different short texts $\langle d_1, d_2 \rangle$, notwithstanding the modules of $RL_{(1)Top\text{-}k}$ and $RL_{(2)Top\text{-}k}$ is the same (i.e. the number of the feature concepts k), the feature concept in the corresponding dimension of the two vectors are different in most case. Without loss of generality, we have to extend each of two relevant concept vectors from original to their union. This is also the reason that the traditional vector model-based information processing methods must establish a high-dimensional vector space which contains all Wikipedia concepts.

Obviously, there are two problems with these vector model-based algorithms: one is that many feature concepts with low (even none) relevance for the target short text participate in computational process, which may cause unexpected side-effect of the algorithm. On the other end of the spectrum, because Wikipedia corpus is very large, one given stem usually only appears at a small number of articles in Wikipedia. So for the target short text, a high-dimensional sparse vector space will be generated. The computation of Cosine over these 0 value dimensional increases the complexity but has no meaning.

From Example 1 we can easily see that, in WordSim-353, the relatedness between "*car*" and "*automobile*" is 8.94 points evaluated by human judgment. Under the case of 10-point system, the score of 8.94 implies that these two terms have rather high semantic relatedness which is consistent with human intuition. However, as we have seen in Table 1, there is only one common concept, "*Car dealership*", between $RL_{(car)Top\text{-}10}$ and $RL_{(automobile)Top\text{-}10}$. So, notwithstanding the modules of the two relevant concept vectors are equal, we cannot use Cosine metric over different dimensions. Consequently, to satisfy the conditions,

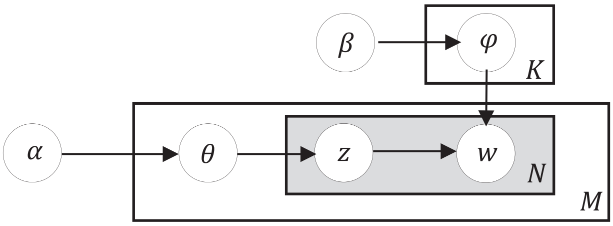


Fig. 3. Graphical model representation of LDA.

the proper vector space for $\langle car, automobile\rangle$ over $RL_{Top\text{-}10}$ should be expressed as $V = RL_{(car)Top\text{-}10} \cup RL_{(automobile)Top\text{-}10}$ (i.e. $|V| = 19$). It is thus easy to perceive that there is only one valid computation with nonzero-value over the dimension named as "*Car dealership*". Along this line, the semantic relatedness assessment returned by Cosine metric will be fairly low.

To cope with the issues mentioned above, this section will study how to assess the semantic relatedness for short text in a relatively low-dimensional vector space without any extension of $RL_{Top\text{-}k}$. Firstly, for two $RL_{(1)Top\text{-}k}$ and $RL_{(2)Top\text{-}k}$, the association between different components in the two vector spaces need to be analyzed. Therefore, we have the following definitions:

Definition 3 (*Association Coefficient for* RL_{Top-k}**).** Given a short text pair $\langle d_1, d_2\rangle$, let $L_1 = \langle C'_1, \dots, C'_k\rangle$ and $L_2 = \langle C''_1, \dots, C''_k\rangle$ be the $RL_{Top\text{-}k}$ defined in Definition 1 for d_1 and d_2 respectively. The *association coefficient* between L_1 and L_2, denoted as $AC_{RL_{Top\text{-}k}}$, can be defined as a k-dimensional vector $AC_{RL_{Top\text{-}k}} = \langle \lambda_1, \dots, \lambda_k\rangle$, where $\lambda_i \in [0,1]$ represents the proximity between Con'_i in C'_i and Con''_i in C''_i, $i \in \{1, \dots, k\}$.

In what follows, we will compute $\lambda_i \in [0,1]$ in $AC_{RL_{Top\text{-}k}}$ by analyzing the topic features over two different vector spaces between $RL_{(1)Top\text{-}k}$ and $RL_{(2)Top\text{-}k}$ for $\langle d_1, d_2\rangle$ without extending the length of dimensions to their union set.

As mentioned in Section 2, the currently common topical feature model is LDA. Therefore, we applies LDA to compute $AC_{RL_{Top\text{-}k}}$ between $RL_{(1)Top\text{-}k}$ and $RL_{(2)Top\text{-}k}$.

As we all know, LDA is a probability-based topic generation model. Multiple distributions of terms and topics contained in datasets can be obtained by unsupervised learning. The smoothed version of LDA model is shown in Fig. 3.

In Fig. 3, M is the total number of articles in the training dataset. N is the total number of terms in article. φ is the distribution of term for topic. K is the total number of topics. θ is the distribution of topic for article. z is the selected topic when a document term is generated. Since one article has multiple topics, the gray box in Fig. 3 indicates that the step of selecting term w and its related topic z is repeated N times. α and β are 2 hyperparameters, and represents the prior Dirichlet distribution of the topic distribution of each article as well as the term distribution of each topic respectively. More details about LDA are introduced in Blei et al. (2003).

According to Definitions 1 and 2, each Wikipedia concept Con_i in $RL_{Top\text{-}k}$ has a unique corresponding Wikipedia page and article. For consistency, we still use Wikipedia as external knowledge source to train LDA model.

Remark 1. It should be noted that according to the definition of information entropy, the terms with too many or too few occurrences in different texts have low distinction. Therefore, in order to reduce the learning cost of LDA, we set a filter window to remove some of these terms.

However, the size of the filter window cannot be set at will. If the value of window size is too small, the filtering effect is not obvious and the learning cost cannot be reduced effectively. On the contrary, if the value is too large, many useful terms will be filtered out, which will affect the accuracy. Based on the experimental results in our previous study (Xiao et al., 2017), the terms those appear less than 20 times in different articles and more than 10% of the total number of articles are removed. The determination of these two values, 20 and 10%, is due to the observation of the relation between the values and the efficiency of the model. In this way, we can get a relatively reasonable balance between accuracy and response efficiency during the training process.

On this basis, by using the Gibbs sampling, two probability matrices φ (term → topic) and θ (topic → article) are obtained under iterative strategy, where φ represents the probability of occurrence of each term in each topic, θ represents the probability of occurrence of each topic in each Wikipedia article. The two probability matrices are formalized as follows:

$$\varphi = \begin{bmatrix} \varphi_{1,1} & \cdots & \varphi_{1,V} \\ \vdots & \ddots & \vdots \\ \varphi_{K,1} & \cdots & \varphi_{K,V} \end{bmatrix} \tag{2}$$

$$\theta = \begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,K} \\ \vdots & \ddots & \vdots \\ \theta_{M,1} & \cdots & \theta_{M,K} \end{bmatrix} \tag{3}$$

where each element in the above two matrices are computed using formulas (4) and (5):

$$\varphi_{k,v} = \frac{n_{k,v} + \beta}{\sum_{i=1}^{V} n_{k,i} + V\beta} \tag{4}$$

$$\theta_{m,k} = \frac{n_{n,k} + \alpha}{\sum_{i=1}^{K} n_{m,i} + K\alpha} \tag{5}$$

Now, we will provide an example to show the fragment of φ (i.e. the term-topic matrix) when using Wikipedia as external knowledge source to train LDA model. By setting the number of topics $K = 3000$, Table 2 display the highest probability of occurrence of the top 8 terms and their corresponding probabilities for 3 topics with the IDs: 100, 103 and 108 respectively. The value in parentheses indicate the probability of term appearing under the target topic.

When the training of the LDA model is completed, according to Definition 3, for a given short text pair $\langle d_1, d_2\rangle$, each $\lambda_i \in AC_{RL_{Top\text{-}k}}$ between $RL_{(1)Top\text{-}k}$ and $RL_{(2)Top\text{-}k}$ can be defined by the following formula:

$$\lambda_i = AssCoe(Con'_i, Con''_i) \tag{6}$$

where $\lambda_i \in [0,1]$ ($i \in \{1, \dots, k\}$), the function $AssCoe(Con'_i, Con''_i)$ denotes the association coefficient between Con'_i and Con''_i at the corresponding position of $RL_{(1)Top\text{-}k}$ and $RL_{(2)Top\text{-}k}$.

Remark 2. It should be noted that both $RL_{(1)Top\text{-}k}$ and $RL_{(2)Top\text{-}k}$ are ordered lists. All elements in both lists are sorted by their tf–idf weights towards the respective d_i ($i \in \{1, 2\}$). Therefore, the topper one element ranks, the more relevant the corresponding concept towards d_i is. Intuitively, it is a reasonable choice for us to compute λ_i just using the elements at the same position in both $RL_{(1)Top\text{-}k}$ and $RL_{(2)Top\text{-}k}$.

It can be seen from Fig. 1 that Wikipedia concept *Con* and Wikipedia article *Art* are one-to-one correspondence. Therefore, $AssCoe(Con'_i, Con''_i)$ can be expressed using the following formula:

$$AssCoe(Con'_i, Con''_i) = AssCoe(Art'_i, Art''_i) \tag{7}$$

where Art'_i and Art''_i denote the corresponding Wikipedia articles of Con'_i and Con''_i respectively.

Along this line, we can map Art'_i and Art''_i to the 2-dimensional topic vector space through LDA model which has been trained, denoted as $\vec{v'}_i$ and $\vec{v''}_i$ respectively. Since values of the topic vector

Table 2
The example of topic-term matrix ($K = 3000$).

Topic ID	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8
100	Scenes (0.375)	Camera (0.280)	Trademark (0.146)	Couples (0.121)	Harrington (0.038)	Faire (0.017)	Detractors (0.013)	Esque (0.011)
103	Subjected (0.229)	Filling (0.221)	Sheila (0.155)	Puppets (0.111)	Condensed (0.084)	Acidic (0.077)	Pigments (0.068)	Alister (0.017)
108	Committed (0.166)	Injured (0.164)	Crimes (0.142)	Convicted (0.105)	Jail (0.074)	Murders (0.064)	Imprisonment (0.061)	Sentences (0.048)

space generated by LDA are probability data, it is necessary to use the metric for probability data to evaluate the difference between the two probability vectors. The current common metrics for measuring probability distribution are Kullback–Leibler divergence (KL), Jensen–Shannon divergence (JS), and Bhattacharyya distance etc., while KL divergence and JS divergence are more widely used. What is more, JS divergence is a deformation of KL divergence, which solves the asymmetry and unboundedness of KL divergence. Therefore, in this paper, we use JS divergence to compute $AssCoe(Art'_i,\ Art''_i)$ given in formula (7). The formula of JS divergence is shown as follows:

$$JSD\left(\vec{v'}_\imath \parallel \vec{v''}_\imath\right) = \frac{1}{2}KLD\left(\vec{v'}_\imath \parallel \frac{\vec{v'}_\imath + \vec{v''}_\imath}{2}\right) + \frac{1}{2}KLD\left(\vec{v''}_\imath \parallel \frac{\vec{v'}_\imath + \vec{v''}_\imath}{2}\right) \tag{8}$$

where $KLD(\vec{v'}_\imath||\vec{v''}_\imath)$ is the KL divergence as shown in formula (9):

$$KLD(\vec{v'}_\imath \parallel \vec{v''}_i) = \sum_{j=1}^{K} v'_{ij} \ln \frac{v'_{ij}}{v''_{ij}} \tag{9}$$

where K is the total number of topics, $v'_{ij} \in \vec{v'}_\imath$ and $v''_{ij} \in \vec{v''}_\imath$.

From formula (8) we can see that, the range of JS divergence is [0, 1], and if P is the same as Q, the result is 0. On the contrary, the result is 1. Therefore, formula (8) needs to be converted to obtain a reasonable result of $AssCoe(Art'_i,\ Art''_i)$. The specific formula is as follows:

$$\begin{aligned} AssCoe(Art'_i, Art''_i) &= 1 - JSD\left(\vec{v'}_\imath \parallel \vec{v''}_\imath\right) \\ &= 1 - \frac{1}{2}(KLD\left(\vec{v'}_\imath \parallel \frac{\vec{v'}_\imath + \vec{v''}_\imath}{2}\right) \\ &\quad + KLD\left(\vec{v''}_\imath \parallel \frac{\vec{v'}_\imath + \vec{v''}_\imath}{2}\right)) \end{aligned} \tag{10}$$

On this basis, the value of λ_i as well as $AC_{RL_{Top\text{-}k}}$ can be obtained by formula (6) to formula (10). For a given short text pair $\langle d_1, d_2\rangle$, in most case, if $d_1 \neq d_2$, then $RL_{(1)Top\text{-}k} \neq RL_{(2)Top\text{-}k}$. For these two different vectors, we can still use $AC_{RL_{Top\text{-}k}}$ to convert the feature concept in $RL_{(2)Top\text{-}k}$ into the following form without using union operation:

$$RL_{(2)Top\text{-}k} = \langle C''_1, \ldots, C''_k\rangle = \vec{\lambda} \cdot RL_{(1)Top\text{-}k} = \langle \lambda_1 \cdot C'_1, \ldots, \lambda_k \cdot C'_k\rangle \tag{11}$$

where $C''_i = \lambda_i \cdot C'_i =$ indicates the degree of association between feature concepts Con'_i and Con''_i at the same component positions of $RL_{(1)Top\text{-}k}$ and $RL_{(2)Top\text{-}k}$.

As a result of the above procedure, we can define a new semantic relatedness computation method over $RL_{(1)Top\text{-}k}$ and $RL_{(2)Top\text{-}k}$ for a given short text pair $\langle d_1, d_2\rangle$ as:

$$\begin{aligned} Rel(d_1, d_2) &= \frac{(\vec{V}(RL_{(1)Top\text{-}k})) \cdot \vec{V}(RL_{(2)Top\text{-}k})}{|\vec{V}(RL_{(1)Top\text{-}k})| \cdot |\vec{V}(RL_{(2)Top\text{-}k})|} \\ &= \frac{\sum_{(C'_i \in RL_{(1)Top\text{-}k} \text{ and } C''_i \in RL_{(2)Top\text{-}k})} \lambda_i \cdot w^{(d_1)}{}_{(C'_i)} \cdot w^{(d_2)}{}_{(C''_i)}}{\sqrt{\sum_{C'_i \in RL_{(1)Top\text{-}k}} \left(w^{(d_1)}{}_{(C'_i)}\right)^2} \cdot \sqrt{\sum_{C''_i \in RL_{(2)Top\text{-}k}} \left(w^{(d_2)}{}_{(C''_i)}\right)^2}} \end{aligned} \tag{12}$$

4.3. The sub-experiment and evaluation for semantic relatedness assessment

On the basis of the above two sub-sections in Section 4, a lower dimensional semi-explicit semantic space can be constructed. Along this line, the semantic relatedness between user queries and the target short text can be computed and sorted to achieve short text retrieval. Obviously, although the computation of semantic relatedness is not the ultimate goal of this paper, it is the key technology in our short text retrieval method. Therefore, we design a sub-experiment on this part specially to further illustrate the effect of semantic relatedness.

Table 3
The benchmarks.

Name	Language	Score range	Number of pairs
Rel-122	English	[0, 4]	122
MTurk-287	English	[0, 5]	287
WordSim-353	English	[0, 10]	353
RW-2034	English	[0, 10]	2034

Noteworthy, semantic relatedness and information retrieval have different evaluation criteria, and most of the researches on semantic relatedness are focused on terms or concepts. So, in this paper we use part of new datasets cited in our previous study (Li et al., 2017b) to elaborate the semantic relatedness experiments.

In this section, the semantic relatedness experiments are made over 4 widely used benchmarks, namely Rel-122 (Szumlanski et al., 2013), MTurk-287.[2] Radinsky et al. (2011), WordSim-353 (Finkelstein et al., 2002), and RW-2034 (Luong et al., 2013). The overview of these benchmarks is shown in Table 3.

It is well known that the evaluation of the accuracy of semantic relatedness is a difficult task because everyone has different subjective attitudes towards relatedness even for the same object. To the best of our knowledge, semantic relatedness assessment can be evaluated using two different correlation coefficients as follows:

(1) Pearson product-moment correlation coefficient (Salton and McGill, 1986), denoted as P. The corresponding formula is defined as:

$$P = \frac{n(\sum_{i=1}^{n} x_i y_i) - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n(\sum_{i=1}^{n} {x_i}^2) - (\sum_{i=1}^{n} x_i)^2}\sqrt{n(\sum_{i=1}^{n} {y_i}^2) - (\sum_{i=1}^{n} y_i)^2}}, \tag{13}$$

where x_i refers to the value of the ith word pair in the dataset given by human judgments, y_i to the corresponding value returned by a certain method, and n to the length of the target dataset.

P reflects the linear correlation between measuring result with human judgments, where 0 means uncorrelated and 1 means perfect correlated.

(2) Spearman rank-order correlation coefficient (Spearman, 1987), denoted as S. The corresponding formula is defined as:

$$S = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n^2 - 1)}, \tag{14}$$

where d_i is the difference between the ranks of x_i and y_i, while the meaning of x_i, y_i and n is the same as above.

S compares the correlation between measuring result with human judgments based on the ranking strategy, where 0 means uncorrelated and 1 means perfect correlated.

Table 4 summarizes the experiment results about Pearson coefficients (P) and Spearman coefficients (S) on different benchmarks given

[2] http://tx.technion.ac.il/kirar/files/Mtruk.csv.

Table 4
Results on Pearson coefficients (P) and Spearman coefficients (S) between related studies.

Related studies	Benchmarks								Reference
	Rel-122		MTurk-287		WordSim-353		RW-2034		
	P	S	P	S	P	S	P	S	
Hadj Taieb et al.	–	–	–	**0.654**	–	0.724	–	–	Hadj Taieb et al. (2013)
Jiang et al.	–	–	–	–	–	0.496	–	–	Jiang et al. (2015)
IC-based (with Lin formula)	–	–	0.5	0.49	0.5	0.48	0.49	0.55	Ben Aouicha et al. (2016b) and Lin (1998)
LWCR (tf–icf)	0.63	0.62	**0.68**	0.58	**0.75**	0.77	–	–	Ben Aouicha et al. (2016a)
LWCR (tf–idf)	0.412	0.387	0.568	0.485	0.601	0.338			Ben Aouicha et al. (2016a)
Li et al.	**0.637**	**0.652**	0.504	0.575	0.555	0.757	**0.523**	**0.621**	Li et al. (2017b)
Our method (k = 10 000)	0.593	0.631	0.537	0.586	0.622	**0.775**	0.494	0.562	

in Table 3, which are obtained from our method as well as some other related studies. The maximum values in each column are marked in bold.

Here, we must note that, based on our experimental analysis in our previous study (Li et al., 2017b), we still set the value of parameter k to 10 000 in this paper.

From the above experimental results shown in Table 4 we know that although some values returned by our method are not very prominent, our method proposed in this paper can obtain the top 3 result on each column over all above four benchmarks. This conclusion reflects the versatility and robustness of our method.

What is more, when we move our eyes to the last two rows of Table 4, our previous study have returned more optimal results. However, the method designed in our previous study uses more iterative strategies which makes the algorithm more complex. Considering that this paper focus on the research about short text retrieval, not semantic relatedness, and the gaps between the optimal results and those returned by the method proposed this paper are very slight, so we still use the semantic relatedness assessment method designed in Section 4 to realize the following short text retrieval. In this way, we can ensure the response efficiency of our method while maintaining certain accuracy.

5. Semi-explicit short text understanding and retrieval model

From the perspective of information retrieval, the query information input by users may be a keyword, a phrase or a sentence. However, since the length of query information is usually not very long, the user's query can also be regarded as a short text. Obviously, the query information and the target short text can be considered as two variables in the formula (12) (i.e. d_1 and d_2) to compute the semantic relatedness between them. Thus, we can sort the search results according to the value of semantic relatedness and return the sorted short text list to the user.

To sum up, by introducing the above semi-explicit semantic features combining Wikipedia concept as well as the corresponding topic information mentioned in Section 4, in this section, we construct the architecture of a new short text understanding and retrieval model as being illustrated in Fig. 4.

6. Experiments and evaluation

For evaluating our new semi-explicit short text retrieval method, in this section we firstly introduce the version of Wikipedia and standard benchmark employed in our experiments. Meanwhile, we use several widely popular standard evaluation metrics which have been applied in many studies about information retrieval. Lastly, we discuss and analyze the experimental results.

6.1. Experimental datasets

In this paper, we implemented our semi-explicit short text retrieval method using a Wikipedia snapshot as of May 20, 2019. The resource can be freely available for download from the link provided by Wikipedia.[3] Besides, our method implements a pre-processing based on Java Wikipedia Library (JWPL)[4] in order to remove some meaningless files which include the lab such as File, Help, Draft, etc. and extract semantic features and depictions for each Wikipedia page. JWPL is an open-source, Java-based application programming interface that allows to access all information contained in Wikipedia. This high-performance Wikipedia API provides structured access to information nuggets like redirects, categories, articles and link structure. After that, we use Java with JavaTM 2 SDK and MySQL to implement our retrieval model given in Section 5.

The standard benchmark used in the experiment is collected from a subset of Twitter in 2011 which contains 39,800,61 user attribute parameters and social network data between users (Kalloubi et al., 2016; Li et al., 2012). The subset collected at least 600 articles from each user, making a total of 500 million English blog with various contents. Moreover, in order to analyze and compare with other related researches, this paper introduces 50 user queries designed in Kalloubi et al. (2016) which have been divided into 20 short text queries and 30 long text queries. The concrete contents of these 50 user queries are shown in Table 5.

The above 50 user queries given in Table 5 are manually selected and judged by an expert using pooling techniques (Manning et al., 2008). The 20 short queries are highlighted in bold. From Table 5, we can conclude that the counts of long text queries and short text queries are divided mainly according to the length of query statements.

Remark 3. As we all know, the data preprocessing task is very important in the field of data science. Too much noise and meaningless page tags can have serious impact on the results. Considering the short length, little information, sparse features and irregular grammar of short text, it is more important to filter and purify the noise. Thus, just as the steps given in the dashed box of Algorithm 1, to improve the accuracy and efficiency of our retrieval model, for both pages as well as the standard benchmark, we firstly build a stop-words list to filter those words which do not have any contribution in semantics. Then the algorithm adopts porter stemming method (Porter, 1980) to transform inflectional forms or derivative forms of a word into a normal form.

What is more, although the content of short texts often do not follow the formal grammar, they always contain key nouns or phrases to express the core meaning. So, in preprocessing stage, we introduce Shallow Parser (Hu et al., 2009) to detect the key phrases in short text and divide the short text into a series of words that together compose a grammatical unit. Here, we must note that, different from POS Tagging, in most case we just need to know a bunch of words together form a noun phrase but do not care about the sub-structure of the tree within other words (e.g. pronoun, determiners, etc.) and how do they combine especially for short text.

Along this line, inspired by the research introduced in Li et al. (2018a), at the stage of implicit topic feature extraction (Section 4.2), we introduce common semantics topic model (CSTM) and create a new type of topic, namely common topic, to gather the noise words in order to further filter the noise information in short text.

[3] https://dumps.wikimedia.org/enwiki/20190520/enwiki-20190520-pages-articles-multistream.xml.bz2.

[4] https://www.ukp.tu-darmstadt.de/software/jwpl.

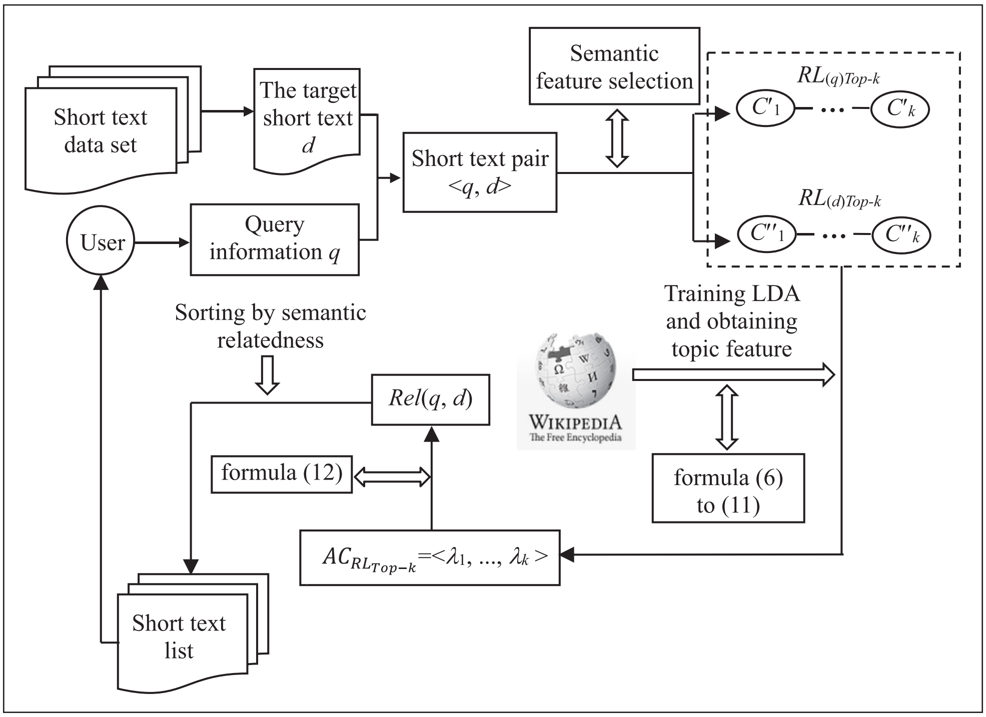


Fig. 4. The architecture of semi-explicit short text understanding and retrieval model.

Table 5
The list of 50 user queries.

Query ID	Query content	Query ID	Query content
	Query ID: (Q1–Q25)		Query ID: (Q26–Q50)
Q1	The social media and the customer relationship in a company	Q26	NHL Western Conference Finals
Q2	Egyptian plane makes unscheduled Landing in Athens due to possible Bomb Threat	Q27	Oprah winfrey ends her twenty five year run of The Oprah Winfrey Show
Q3	Barack Obama announces that Osama bin Laden was killed in a military operation	Q28	**Indianapolis 500: Dan Wheldo**
Q4	The wikileaks twitter account	Q29	**Indianapolis**
Q5	The online events academy	Q30	Obama rejects Bush Iraq withdrawal plan
Q6	Perspective of social media in education	Q31	Iraq Oil Industry
Q7	Captain's share by Nathan Lowell	Q32	UK Firms Invited to Explore Opportunities in Iraq
Q8	LinkedIn starts social news services	Q33	Security Council extends UN mission in Iraq
Q9	Recommendation for a new bank and wamu services	Q34	Google Stays in China And Baidu Keeps on Winning
Q10	**Google buzz**	Q35	The Turkey Kurdish Conflict
Q11	**Osama Bin laden**	Q36	The political structure of Turkey
Q12	**United States Army**	Q37	Groupon Launches iPhone Application in the UK
Q13	**Captain America**	Q38	**Sony**
Q14	What is new in the world of Apple? Samsung?	Q39	**Tour de France**
Q15	**Football team**	Q40	**Lance Armstrong**
Q16	The president Barak Obama and his speech about Osama Bin Laden	Q41	Thor Hushovd just won another stage in Tour de france
Q17	**The U.S department of labor**	Q42	**BBC News**
Q18	**The U.S justice department**	Q43	Explosion hits government offices in Oslo
Q19	**Pirates of the Caribbean**	Q44	A government building has been damaged in Norway's capital
Q20	Which films wins the Palme d'Or in Cannes Film Festival?	Q45	BlackBerry PlayBook Available in the US and Canada
Q21	**The Tree of Life**	Q46	**The FIFA world cup**
Q22	**The Billboard Music Awards**	Q47	**The BMW cars**
Q23	**The Country Music**	Q48	Bin Hammam accused of buying 2022 World Cup
Q24	**UEFA Champions League**	Q49	Egyptians fired up their revolt against Hosni Mubarak
Q25	Barcelona beats Manchester United and wins Champions League final	Q50	Egypt tells Iran to stay out of Arab's business

6.2. Evaluation metrics

Considering the introduction of the ranking strategy in retrieval process, the three well-known metrics, Mean Average Precision (*MAP*), Precision at rank k (*P@k*) and *R-Prec*, wildly applied in current information retrieval methods are used as evaluation metrics to measure the validity of the proposed short text retrieval method. More details about these 3 metrics are introduced in Kalloubi et al. (2016). The corresponding formulas are defined as follows:

(1) *MAP* refers to the mean of the average precision rate in all queries:

$$MAP = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{Q_j}\sum_{i=1}^{Q_j}P(T_i) \tag{15}$$

where N is the total number of queries, Q_j is the number of relevant documents returned by query j, and $P(T_i)$ is the precision rate of the relevant document at the location where the document is returned. *MAP* can directly reflect the single-valued indicator for the performance of the system over all relevant documents. The topper location one document is returned, the higher score it should be ranked.

(2) *P@k* refers to the precision rate of the top k results returned by a given retrieval method (in this paper, we set $k = 30$):

$$P@k = \frac{\text{The number of relevant documents in top } k \text{ results for a given query}}{k} \tag{16}$$

(3) *R-Prec* refers to the precision rate when R documents are retrieved, where R is the total number of relevant documents for the given query. The system returns R documents, among which there are r relevant documents in total:

$$R - Prec = \frac{r}{R} \tag{17}$$

6.3. Evaluation and analysis

Referring to our previous research conclusions (Li et al., 2017b; Xiao et al., 2017), Table 6 gives the values of relevant parameters for better experimental results.

As mentioned in Section 2, different strategies as well as experimental datasets are being used in different studies about short text understanding and retrieval. Obviously, it is difficult to directly compare and evaluate the experimental results shown in different studies. The apples-vs.-oranges comparison between experimentations on different datasets is meaningless. Therefore, in order to ensure the consistency of the experimental data and analyze the experimental results more clearly, some short text retrieval methods with high correlation presented in the past few years are selected for comparative experiments. Moreover, two classic models, ESA and LDA, are also applied over the standard benchmark to further illustrate the effectiveness of the proposed method. By using formulas (15) to (17), Table 7 summarizes the results of several methods on the standard benchmark given in Section 6.1. The corresponding experimental results shown in rows 3 and 4 of Table 7 are referred from Kalloubi et al. (2016) and Xiao et al. (2017), respectively.

Making the longitudinal comparison between the results of the columns in Table 7, we can conclude that the short text retrieval method proposed in this paper improves the retrieval results over each evaluation metrics in both two kinds of query requires when comparing with several other retrieval methods. Obviously, by selecting feature concepts in Wikipedia and analyzing the topic information of the feature concept, we can discover more potential semantics and construct a more reasonable short text understanding and retrieval model.

What is more, when making the latitudinal comparison between two kinds of query requirements, it can be seen that the performance of long text queries is better than those of short text queries over all the five retrieval methods listed in Table 7. This is because that, from the view of the length of text, long text always contains more information. The original semantic information carried by the text itself is often more accurate than the information extended artificially, and has better semantic connection between contexts. Therefore, the length of the text has a very important influence on the retrieval results.

On the above basis, we make the further comprehensive statistics and analyses over all 50 query results for above several retrieval methods on the standard benchmark. The specific results are shown in Table 8, and the experimental results of the "Baseline system" is also from Kalloubi et al. (2016). Concretely, the baseline system uses the degree centrality factor, which has been shown to obtain optimal results in word sense disambiguation (Navigli and Lapata, 2010). We can observe from Table 8 that by taking Wikipedia concept feature and topic information into consideration, the semi-explicit short text retrieval method proposed in this paper can improve the retrieval results over each evaluation metric.

To be clear, in fact we can also set $k = 10$ or 20 for *P@k* during the experiment of this paper. The reason why we set $k = 30$ in the metric, *P@k* (see Tables 7 and 8) is that in some existing related researches (Kalloubi et al., 2016; Xiao et al., 2017), the authors have used $k = 30$ to evaluate the retrieval effect. Just as mentioned in Kalloubi et al. (2016), "*P*@30 was used as the official measurement in the TREC Microblog ad-hoc task". So, if we assign some other values to k, there is no corresponding experimental data in aforementioned related researches. Therefore, in order to be able to directly refer the experimental results in other existing related researches for comparison, we also set $k = 30$, so as to more clearly illustrate the effectiveness of our method.

Noteworthy, from Tables 6 to 8, we can see that for a given standard benchmark, although only a few Wikipedia concepts ($k = 10\,000$) are used to construct vector space (about 0.18% of dimensions of the whole vector space), the short text retrieval method proposed in this paper can return pretty competitive evaluation results on *MAP*, *P@k* and *R-Prec*.

Moreover, inferred from the sub-experiment mentioned in Section 4.3 as well as our previous research conclusions (Li et al., 2017b), when $k = 10\,000$, the feature selection algorithm (i.e. Algorithm 1) can build a more effective feature space, which can make Pearson and Spearman coefficients relatively converge and get better results for the evaluation of semantic relatedness. Meanwhile, the computational efficiency of the algorithm is also in an acceptable range. Although the research object and implementation process designed in this paper are different from our previous study, the strategy of feature selection is similar, so $k = 10\,000$ is still used in this paper. More importantly, from the perspective of experiments, the results are also satisfactory. Obviously, our method may perform even better with the increase of the threshold k if we are willing to accept higher computational complexity.

On the whole, by filtering the explicit concept features and analyzing implicit topic information, our method can discover more deep semantics both in user queries and target short texts, and extract the most relevant ones. It gives our method the ability to better match the needs of users.

7. Conclusions

This paper presents a methodology to realize short text retrieval by assessing semantic relatedness combining Wikipedia concept feature and the corresponding topic information. To tackle the limitations of traditional information retrieval methods in short text retrieval task, we have mainly undertaken the following works in this paper. Firstly, we design a feature selection algorithm to return the top k most relevant Wikipedia concepts, for a given short text. Thus, based on the LDA model, by analyzing the corresponding topic information of the feature concepts in Wikipedia, we propose some formulas to determine the association coefficient list between different components of the corresponding positions in two $RL_{(i)Top\text{-}k}$ ($i \in \{1, 2\}$), thereby transforming two different feature vectors into the same semantic space. Along this line, using $RL_{Top\text{-}k}$ as the whole vector space, a novel semi-explicit short text retrieval method is presented by assessing and sorting the semantic relatedness between the user query and the target short text in dataset under this lower dimensional semantic space. The evaluation, based on the standard benchmark as well as several widely used metrics, indicates that the proposed method in this paper displays a better performance than some state-of-the-art systems.

As future work, we are planning to do some works on short text classification and clustering by applying the semantic relatedness assessment presented in this paper. In addition, instead of employing Wikipedia, we will use some ontology-based knowledge graph in Linked Data as the target knowledge source to introduce the reasoning

Table 6
The parameter values for the better experimental results.

Algorithm procedure	Parameter name	Parameter value
Feature concept selection	The length of RL_{Top-k} (k)	10 000
Topic feature generation	The number of iterations of Gibbs sampling	1000
	The total number of topics (K)	1000
	α	0.05
	β	0.001

Table 7
Evaluation results of several retrieval methods on different retrieval requirements.

Related retrieval methods	Evaluation over short text queries			Evaluation over long text queries			Reference
	MAP	*P@k* (k = 30)	*R-Prec*	*MAP*	*P@k* (k = 30)	*R-Prec*	
LDA	0.537	0.669	0.513	0.635	0.718	0.607	Blei et al. (2003)
ESA	0.498	0.577	0.49	0.563	0.676	0.552	Gabrilovich and Markovitch (2007)
Kalloubi et al.	0.550	0.695	0.531	0.688	0.764	0.674	Kalloubi et al. (2016)
Xiao et al.	0.553	0.721	0.542	0.733	0.854	0.71	Xiao et al. (2017)
Our method	0.626	0.769	0.605	0.761	0.883	0.732	–

Table 8
The summary evaluation results of several retrieval methods.

Related retrieval methods	Evaluation over all 50 queries			Reference
	MAP	*P@k* (k = 30)	*R-Prec*	
LDA	0.602	0.693	0.595	Blei et al. (2003)
ESA	0.539	0.627	0.52	Gabrilovich and Markovitch (2007)
Baseline system	0.593	0.64	0.582	Kalloubi et al. (2016)
Kalloubi et al.	0.638	0.719	0.601	Kalloubi et al. (2016)
Xiao et al.	0.657	0.765	0.632	Xiao et al. (2017)
Our method	0.736	0.828	0.69	–

mechanism for discovering the deeper semantic in short texts. Last but not least, inspired by the research in Ben Aouicha et al. (2016b), Milne and Witten (2013) and Zhang et al. (2011), we aim to apply the proposed short text retrieval method to some cross-language tasks.

CRediT authorship contribution statement

Pu Li: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Supervision, Validation, Writing - original draft, Writing - review & editing. **Tianci Li:** Data curation, Software, Validation. **Suzhi Zhang:** Resources, Data curation, Project administration, Supervision. **Yuhua Li:** Data curation, Methodology, Software, Validation. **Yong Tang:** Formal analysis, Funding acquisition, Investigation, Supervision. **Yuncheng Jiang:** Conceptualization, Funding acquisition, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The works described in this paper are supported by The National Natural Science Foundation of China under Grant Nos. 61802352, 61772211, 61772210, 61975187; The Project of Science and Technology in Henan Province, China under Grant Nos. 202102210178, 202102210387, 202102210397; The Ph.D. Research Foundation of Zhengzhou University of Light Industry, China under Grant No. 0215/13501050015; The School Program Foundation for Training Young Key Teachers of Zhengzhou University of Light Industry, China under Grant No. 2018XGGJS006.

References

Abel, F., Celik, I., Houben, G.J., Siehndel, P., 2011. Leveraging the semantics of tweets for adaptive faceted search on twitter. In: Proceedings of International Semantic Web Conference. Springer, Berlin, Heidelberg, pp. 1–17.

Alsmadi, I., Gan, K.H., 2019. Review of short-text classification. Int. J. Web Inf. Syst. 15 (2), 155–182.

Azad, H.K., Deepak, A., 2019. Query expansion techniques for information retrieval: A survey. Inf. Process. Manag. 56 (5), 1698–1735.

Bekkali, M., Lachkar, A., 2019. An effective short text conceptualization based on new short text similarity. Soc. Netw. Anal. Min. 9 (1), 1.

Ben Aouicha, M., Hadj Taieb, M.A., Ben Hamadou, A., 2016a. LWCR: multi-layered Wikipedia representation for computing word relatedness. Neurocomputing 216, 816–843.

Ben Aouicha, M., Hadj Taieb, M.A., Ben Hamadou, A., 2016b. Taxonomy-based information content and wordnet-wiktionary-wikipedia glosses for semantic relatedness. Appl. Intell. 45, 1–37.

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., 2003. A neural probabilistic language model. J. Mach. Learn. Res. 3 (2), 1137–1155.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S., 2009. Dbpedia-a crystallization point for the Web of data. J. Web Semant. 7 (3), 154–165.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3 (1), 993–1022.

Chen, J., Nairn, R., Nelson, L., Bernstein, M.S., Chi, E.H., 2010. Short and tweet: experiments on recommending content from information streams. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, Atlanta, Georgia, USA, pp. 1185–1194.

Chen, C.T., Ren, J.T., 2017. An improved PLDA model for short text. In: Proceedings of International Conference on Applications of Natural Language Processing and Information Systems. Springer, Cham, pp. 58–70.

Chen, Y., Zhang, H., Liu, R., Ye, Z.W., Lin, J.Y., 2019. Experimental explorations on short text topic mining between LDA and NMF based Schemes. Knowl.-Based Syst. 163, 1–13.

Chu, T.Z., Cheng, L., Wong, H.S., 2017. Corpus-based topic diffusion for short text clustering. Neurocomputing 275, 2444–2458.

Cuong, H.N., Tran, V.D., Van, L.N., Than, K., 2019. Eliminating overfitting of probabilistic topic models on short and noisy text: the role of dropout. Internat. J. Approx. Reason. 112, 85–104.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., 1990. Indexing by latent semantic analysis. J. Amer. Soc. Inf. Sci. 41 (6), 391–407.

Ensan, F., Al-Obeidat, F., 2019. Relevance-based entity selection for ad hoc retrieval. Inf. Process. Manage. 56 (5), 1645–1666.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E., 2002. Placing search in context: the concept revisited. ACM Trans. Inf. Syst. 20, 116–131.

Gabrilovich, E., Markovitch, S., 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of International Joint Conference on Artifical Intelligence (IJCAI 2007). Morgan Kaufmann, Hyderabad, India, pp. 1606–1611.

Gan, C., Wang, W., 2015. Uses and gratifications of social media: a comparison of microblog and WeChat. J. Syst. Inf. Technol. 17 (4), 351–363.

Hadj Taieb, M.A., Ben Aouicha, M., Ben Hamadou, A., 2013. Computing semantic relatedness using Wikipedia features. Knowl.-Based Syst. 50, 260–278.

Han, Z.Y., Yang, M.Y., Kong, L.L., Qi, H.L., Li, S., 2016. Query expansion based on term time distribution for microblog retrieval. Chinese J. Comput. 39 (10), 2031–2044.

Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G., 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence 194, 28–61.

Hofmann, T., 2004. Probabilistic latent semantic indexing. In: Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, Sheffield, U.K, pp. 56–73.

Hu, X., Sun, N., Zhang, C., Chua, T.S., 2009. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management. ACM, Hong Kong, China, pp. 919–928.

Huang, P.S., Chiu, P.S., Chang, J.W., Huang, Y.M., Lee, M.C., 2019. A study of using Syntactic Cues in Short-text Similarity Measure. J. Internet Technol. 20 (3), 839–850.

Huang, H., Wang, Y., Chong, F., Liu, Z., Qiang, Z., 2017. Leveraging conceptualization for short-text embedding. IEEE Trans. Knowl. Data Eng. 30 (7), 1282–1295.

Jiang, Y., Zhang, X., Tang, Y., Nie, R., 2015. Feature-based approaches to semantic similarity assessment of concepts using Wikipedia. Inf. Process. Manage. 51, 215–234.

Kalloubi, F., Nfaoui, E.H., Beqqali, O.E., 2016. Microblog semantic context retrieval system based on linked open data and graph-based theory. Expert Syst. Appl. 53, 138–148.

Ke, T., Abel, F., Hauff, C., Houben, G.J., 2012. Twinder: a search engine for twitter streams. In: Proceedings of International Conference on Web Engineering. Springer, Berlin, Heidelberg, pp. 153–168.

Kozlowski, M., Rybinski, H., 2019. Clustering of semantically enriched short texts. J. Intell. Inf. Syst. 53 (1), 69–92.

Lau, C.H., Tao, X., Tjondronegoro, D., Li, Y., 2012. Retrieving information from microblog using pattern mining and relevance feedback. In: Proceedings of International Conference on Data and Knowledge Engineering. Springer, Berlin, Heidelberg, pp. 152–160.

Le, Q.V., Mikolov, T., 2014. Distributed representations of sentences and documents. Comput. Sci. 4, 1188–1196.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Kleef, P.V., Auer, S., 2015. Dbpedia-a large-scale, multilingual knowledge base extracted from Wikipedia. Semant. Web 6 (2), 167–195.

Li, R., Wang, S., Deng, H., Wang, R., Chang, C.C., 2012. Towards social user profiling: unified and discriminative influence model for inferring home locations. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Beijing, China, pp. 1023–1031.

Li, P., Xiao, B., Ma, W.J., Jiang, Y.C., Zhang, Z.F., 2017b. A graph-based semantic relatedness assessment method combining Wikipedia features. Eng. Appl. Artif. Intell. 65, 268–281.

Li, J., Yi, C., Cai, Z., Leung, H., Kai, Y., 2017a. Wikipedia based short text classification method. In: Proceedings of International Conference on Database Systems for Advanced Applications. Springer, Cham, pp. 275–286.

Li, X., Yue, W., Zhang, A., Li, C., Chi, J., Ouyang, J., 2018a. Filtering out the noise in short text topic modeling. Inform. Sci. 456, 83–96.

Li, X., Zhang, A., Li, C., Guo, L., Wang, W., Ouyang, J., 2018b. Relational biterm topic model: short-text topic modeling using word embeddings. Comput. J. 62 (3), 359–372.

Liang, S., Ren, Z., Rijke, M.D., 2014. The impact of Semantic Document Expansion on Cluster-Based Fusion for Microblog search. In: Proceedings of European Conference on Information Retrieval. Springer, Cham, pp. 493–499.

Lin, D., 1998. An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning. Madison, Wisconsin, USA, July, pp. 296–304.

Liu, D., Fu, Q., Wei, Y., Wan, C., Liu, X., Zhong, M., Qiu, J., 2018. Social short text retrieval based on multiple-enhanced graph and topic model. J. Chinese Inf. Process. 32 (3), 110–119.

Liu, W., Quan, X., Feng, M., Qiu, B., 2010. A short text modeling method combining semantic and statistical information. Inform. Sci. 180 (20), 4031–4041.

Lu, K., Roa, D., Fang, H., 2014. Concept based tie-breaking and maximal marginal relevance retrieval in microblog retrieval. In: Proceedings of the 23rd Text Retrieval Conference. Gaithersburg, Maryland, USA, pp. 1–4.

Lund, K., Burgess, C., 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. Behav. Res. Methods Instrum. Comput. 28 (2), 203–208.

Luong, T., Socher, R., Manning, C.D., 2013. Better word representations with recursive neural networks for morphology. In: Proceedings of the 17th Conference on Computational Natural Language Learning. Sofia, Bulgaria, August, pp. 104–113.

Mahdisoltani, F., Biega, J., Suchanek, F., 2014. Yago3: A knowledge base from multilingual Wikipedias. In: Proceedings of the 7th Biennial Conference on Innovative Data Systems Research. Asilomar, California, USA, pp. 1–12.

Manning, C.D., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press, New York, USA.

Meij, E., Weerkamp, W., Rijke, M.D., 2012. Adding semantics to microblog posts. In: Proceedings of the 5th ACM International Conference on Web Search and Data Mining. ACM, Seattle, Washington, USA, pp. 563–572.

Mendes, P.N., Passant, A., Kapanipathi, P., Sheth, A.P., 2010. Linked Open Social Signals. In: Proceedings of International Conference on Web Intelligence & Intelligent Agent Technology. IEEE, Toronto, Canada, pp. 224–231.

Mikolov, T., Karafiát, M., Burget, L., Cernocky, J., Khudanpur, S., 2010. Recurrent neural network based language model. In: Proceedings of Conference of the International Speech Communication Association. Makuhari, Chiba, Japan, pp. 1045–1048.

Milne, D., Witten, I.H., 2013. An open-source toolkit for mining Wikipedia. Artificial Intelligence 194, 222–239.

Mohamed, M., Oussalah, M., 2019. SRL-ESA-TextSum: A text summarization approach based on semantic role labeling and explicit semantic analysis. Inf. Process. Manage. 56 (4), 1356–1372.

Nasir, J.A., Varlamis, I., Ishfaq, S., 2019. A knowledge-based semantic framework for query expansion. Inf. Process. Manage. 56 (5), 1605–1617.

Navigli, R., Lapata, M., 2010. An experimental study of graph connectivity for unsupervised word sense disambiguation. IEEE Trans. Pattern Anal. Mach. Intell. 32 (4), 678–692.

Porter, M.F., 1980. An algorithm for suffix stripping. Program 14 (3), 130–137.

Qiang, J., Li, Y., Yuan, Y., Liu, W., Wu, X., 2019. A practical algorithm for solving the sparseness problem of short text clustering. Intell. Data Anal. 23 (3), 701–716.

Qu, R., Fang, Y.Y., Bai, W., Jiang, Y.C., 2018. Computing semantic similarity based on novel models of semantic representation using Wikipedia. Inf. Process. Manage. 54, 1002–1021.

Radinsky, K., Agichtein, E., Gabrilovich, E., Markovitch, S., 2011. A word at a time: computing word relatedness using temporal semantic analysis. In: Proceedings of the International Conference on World Wide Web. Hyderabad, India, March, pp. 337–346.

Salton, G., McGill, M.J., 1986. Introduction to Modern Information Retrieval. McGraw-Hill, New York.

Song, Y.Q., Upadhyay, S., Peng, H.R., Mayhew, S., Roth, D., 2019. Toward any-language zero-shot topic classification of textual documents. Artificial Intelligence 274, 133–150.

Song, Y., Wang, H., Wang, Z., Li, H., Chen, W., 2011. Short text conceptualization using a probabilistic knowledgebase. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI 2011). Morgan Kaufmann, Barcelona, Spain, pp. 2330–2336.

Spearman, C., 1987. The proof and measurement of association between two things. Amer. J. Psychol. 100, 441–471.

Suchanek, F.M., Kasneci, G., Weikum, G., 2008. Yago: A large ontology from Wikipedia and WordNet. Web Semant. Sci. Serv. Agents World Wide Web 6 (3), 203–217.

Szumlanski, S., Gomez, F., Sims, V.K., 2013. A new set of norms for semantic relatedness measures. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria, August, pp. 890–895.

Tajbakhsh, M.S., Bagherzadeh, J., 2019. Semantic knowledge LDA with topic vector for recommending hashtags: Twitter use case. Intell. Data Anal. 23 (3), 609–622.

Tang, J.L., Wang, X.F., Gao, H.J., Hu, X., Liu, H., 2012. Enriching short text representation in microblog for clustering. Front. Comput. Sci. China 6 (1), 88–101.

Vicient, C., Moreno, A., 2015. Unsupervised topic discovery in micro-blogging networks. Expert Syst. Appl. 42 (17–18), 6472–6485.

Wang, Z., Cheng, J., Wang, H., Wen, J., 2016. Short text understanding: a survey. J. Comput. Res. Dev. 53 (2), 262–269.

Wang, Z., Zhao, K., Wang, H., Meng, X., Wen, J.R., 2015. Query understanding through knowledge-based conceptualization. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI 2015). Morgan Kaufmann, Buenos Aires, Argentina, pp. 3264–3270.

Xiao, B., Li, P., Hu, J., Jiang, Y., 2017. Microblog semantic retrieval based on latent semantic and graph structure. Comput. Eng. Des. 43 (6), 182–188, 194.

Yahav, I., Shehory, O., Schwartz, D., 2018. Comments mining with tf-idf: the inherent bias and its removal. IEEE Trans. Knowl. Data Eng. 31 (3), 437–450.

Yang, S., Huang, G., Cai, B., 2019. Discovering topic representative terms for short text clustering. IEEE Access 7, 92037–92047.

Yao, L., Pan, Z., Ning, H., 2018. Unlabeled short text similarity with LSTM encoder. IEEE Access 7, 3430–3437.

Zhang, C., Fan, X., Chen, X., 2011. Hot topic detection on Chinese short text. In: Proceedings International Conference on Computer Education, Simulation and Modeling. Springer, Berlin, Heidelberg, pp. 207–212.

Zhang, H., Zhong, G., 2016. Improving short text classification by learning vector representations of both words and hidden topics. Knowl.-Based Syst. 102, 76–86.

Zhu, L., Xu, H., Xu, Y., Xiao, Y., Li, J., Deng, J., Sun, X., Bai, X., 2019. A joint model of extended LDA and IBTM over streaming Chinese short texts. Intell. Data Anal. 23 (3), 681–699.