World Wide Web
https://doi.org/10.1007/s11280-018-0584-z

Extended search method based on a semantic hashtag graph combining social and conceptual information

Wanqiu Cui[1] · **Junping Du**[1] · **Dawei Wang**[2] · **Feifei Kou**[1] · **Meiyu Liang**[1] · **Zhe Xue**[1] · **Nan Zhou**[1]

Received: 30 November 2017 / Revised: 30 March 2018 / Accepted: 30 April 2018

Abstract Searching for microblog short text by their meaning is a challenging task because of the semantic sparsity of the information in social networks. The extended search approaches are commonly accepted which facilitate short text understanding and search by enriching the short text. However, they only analyze the literal semantics of short text, and the unique social characteristics of social network which also contain semantic information are not utilized well. To better capture the rich semantics in microblog short text, we propose a new microblog short text extended search method based on a semantic hashtag graph by combining social and

This article belongs to the Topical Collection: *Special Issue on Web and Big Data*
Guest Editors: Junjie Yao, Bin Cui, Christian S. Jensen, and Zhe Zhao

✉ Junping Du
junpingdu@126.com

Wanqiu Cui
wanqiu.wd@gmail.com

Dawei Wang
devy.wq@ruc.edu.cn

Feifei Kou
koufeifei000@126.com

Meiyu Liang
meiyu-1210@126.com

Zhe Xue
xuezhe@bupt.edu.cn

Nan Zhou
zhounan345@163.com

[1] Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China

[2] School of Information, Renmin University of China, Beijing, China

Published online: 17 July 2018

conceptual information, which enriches each short text by concepts and associated hashtags to represent whole semantic features. Considering the microblog context, we introduce concepts through Wikipedia, as well as semantic consistency of hashtags. Specifically, for conceptual semantics, we propose a conceptual analysis method which merges explicit and implicit information in Wikipedia. For social semantics in hashtags, a semantic hashtag graph which combines social and conceptual information is put forward to generate semantic associated hashtags. We conduct experiments and the results show that our method is obviously better than the other existing state-of-the-art approaches in semantic understanding and search of short text.

Keywords Extended search · Semantic analysis · Social characteristics · Conceptual semantics · Social semantics

1 Introduction

A wide range of applications embark on short text processing as they are widely disseminated and shared on the social platform in recent years. Its search form is also extended continuously to enhance the user's search experience. As a research hotspot in the search form, the semantic analysis and search of short text are widely studied [17, 18]. Such as the search on user-oriented [19, 24] and implementation of precise search for the short text of microblog [16, 44]. However, the search of short text has always been a challenge because of the sparsity and limitation of semantics. For the short text in microblog, it contains social auxiliary information such as hashtags. This information makes it possible to have more available potential information when analyzing semantics than plain text (i.e. news headlines). Through the investigation of multiple datasets, about 19% of microblogs include hashtags, 10% include mention (@), and 14% contain link information (URL). As a guidance auxiliary information for microblog, the hashtags often categorize text quickly according to the meaning of texts. Most of them are used as the external metadata to describe the text content. If it is fully utilized, the hashtag will make short text to reflect more semantic information in the search. However, the traditional microblog search method usually neglects the analysis of hashtags or treats them equally as texts. The purpose of this paper is to solve the problem of the semantic sparsity of the microblog short text in search. We extend more semantics and topic-directed information (which is a kind of social semantics) to the short text by integrating hashtag-guided topic trends and the relationship between microblog texts based on the conceptual semantics in microblog search.

On the one hand, the semantic analysis and extension have yielded many results in short text understanding and search. The typical topic models [45, 47] analyze the topic of the texts by overcoming the lack of contextual information. The extended search methods include pseudo relevance feedback methods [1, 7, 9] using search results to extend, the concept-based extension methods [4, 23, 34, 42] to get the relative concepts through the external knowledge base as well as the methods of using temporal features [11, 26, 41] and so on. They all proceed from the simple text semantics, without considering the topic consistency of microblog information and the social relations in search. On the other hand, the extension method is also proposed for microblog search [2]. The hashtags [3, 36] and microblog structure information [25] are used to assist search, but they did not consider the text semantics. Therefore, on the basis of related

short text research, we make full use of the characteristics of hashtag and social auxiliary information to combine the text semantics and microblog topic trends. Therefore, incorporating the conceptual and social semantics when searching for short text in microblog may further improve the quality of short text search.

We propose an extended search method which enriches the conceptual semantics and social semantics in hashtags of the short text by integrating the topic consistency and social relations of microblog data. For conceptual semantics acquisition, traditional explicit semantic analysis (ESA) method [10] ignores the structure and potential information of the concept page which leads to the omission of some related concepts. So we design a new E&ISA algorithm which divides the concept page to explicit and implicit information, and mine more of the hidden concepts to generate the concepts closer to the meaning of the text. For the generation of social semantics in hashtags, although most of the information released by microblog is related to a social background and theme, the proportion of hashtag does not cover all the microblog data space [3]. This causes difficulty in constructing hashtag semantics in this paper. To solve this problem, we regard the concepts as the hashtags of plaint text to achieve unified modeling. The model is created by connecting the related hashtags. We define it as a hashtag graph model combining the social and conceptual semantics (SCSHG). The associated hashtags are generated in the conceptual context achieved by E&ISA. Then, we put forward the extended search method based on SCSHG model (SCSHG-ES). According to the extended semantics, BM25F ranking algorithm is used to search. The experimental results show that our method can improve the effectiveness and quality of the short text search in microblog.

The main contributions of our work are summarized as follows:

1. In order to extract more semantic representations of short text in the search, we put forward a new extended search method named SCSHG-ES. It enriches the semantics of short text by concepts and associated hashtags, thereby improving the search performance in the microblog short text.
2. We introduce a combination of explicit and implicit semantic analysis algorithm (E&ISA) based on the Wikipedia to obtain conceptual semantics. It is able to tap the potential of conceptual knowledge in Wikipedia fully, with more relevant semantics as concept extension.
3. We form the unified format of the text of microblog through using the concepts of plaint text as hashtags. It solved the problem that some data (without the hashtags) cannot be fused when building a graph model by hashtags. Moreover, it lays the foundation for extracting the hashtag semantics to extend the short text.
4. We construct the SCSHG model using hashtags and social relationship, which combines conceptual semantics and social features. On the basis of this model, the associated hashtag features of related events are mined deeply, and help to achieve accurate search as a further extension of the textual social semantics.

The rest of the paper is organized as follows. Section 2 provides the related work, about research on semantic understanding and extended search for short text. Sections 3 describes the details of our method, including the overview of the SCSHG-ES method and the implementation process. Section 4 presents a series of experiments, and evaluates the search performance of SCSHG-ES method, and we conclude the whole paper in Section 5.

2 Related work

2.1 Semantic understanding of short texts

Regarding to short text search, earlier works improve the search efficiency from the analysis of text semantics. In particular, the classical LSA [39] and topic model algorithm [33, 43] aim to capture the latent semantics and establish a document relevance beyond the lexical overlap, to a certain extent reducing the semantic gap in short textual representations. They construct a learning model to improve semantic search based on statistics. It is easy to cause the lack of text semantics through ignoring the semantic relationship between texts. Therefore, J sun et al. [32] proposed an interactive strategy to infer the latent query semantics by learning from user feedbacks. It is superior to the probabilistic topic model method, and the text semantics can be fully understood through the finite interaction. However, because of limited number of words contained in the short text, it is a challenge to describe the semantics completely. To deal with these problems, the researchers are widely devoted to the study of semantic analysis [14, 31]. Wang et al. [37] studied the understanding of short text. They analyzed a variety of processing techniques for short text, and the understandable semantic relations are mined in search.

2.2 Extended search for short texts

Due to the shortness and ambiguity of the short text, it is difficult to semantically analyze literally in search. X. Sean et al. [35] proposed the user entity and association relations to achieve intelligent search. In addition, extended search method by using high-quality external knowledge base can enhance the semantic expression ability of text. Increasing the number of feature words can also solve the sparseness problem of short text to a certain extent. The extended methods based on user profiles [46], temporal and spatial characteristics [22] enrich the semantics of short text with related knowledge. The bag of concepts [26, 42] and the feedback conceptual model [38] are used to obtain the representation of the short text with the relevant concepts.

In particular, the concept model is regarded as a popular semantic analysis method for text and image [12] because of its effectiveness in semantic learning. The concept-based search is the use of external knowledge sources to provide additional background knowledge and contexts that are not explicitly appeared in document and queries. For example, ESA [10] and LexSA [16] based on the Wikipedia [8, 15, 20] and Probase [40] knowledge base construct the vector of concepts to express the deeply semantics of short text. Those method is better than the bag of words and topic model, and can provide more reference text semantics.

However, the bag of concepts and concept feedback model only analyze the text. Although a unified knowledge model has been formed on semantic understanding, the meaning of short text does not always obviously in the literal and the user's attention topics will change with time in the microblog search. So we put forward the conceptual representation as the central language, and combine with hashtags and social attributes to achieve extended search. Compared with bag of words and concepts method, the extension of short text by using the semantics of concept and auxiliary information can make search benefit from a large number of external knowledge and social network environment. It can meet the needs of search.

3 Extended search method based on SCSHG model

In this section, we describe the details of the proposed SCSHG-ES method. More specially, this method is implemented in three sections formalizing the issues of the extension of conceptual semantic feature, the extension of social semantic feature in hashtags and extended search for microblog short text respectively. The notations used throughout this paper is shown in Table 1.

3.1 The framework of SCSHG-ES

We propose a new extended search method for short text called SCSHG-ES. Instead of extending the concepts or pseudo relevance feedback information like existing approaches do, we extend the microblog short text with the concepts and the hashtag information. We assume to extend the short text becoming a virtual structure which contains three fields, which are the original text, the conceptual semantic feature and the social semantic feature in hashtag, represented as *ST*, *CS*, and *HS* respectively. The short texts of microblog consist of the hashtag and plaint text, which is collectively called short text in this paper, denoted by $ST = \{ST_1, \ldots, ST_i\}$, and $|ST_i| >= 1$. The *CS* are generated after the conceptual extended learning of *ST* using E&ISA algorithm. The *HS* are generated by the SCSHG model, and finally the semantic feature structure of *ST*` are represented by Eq. (1):

$$ST` \rightarrow \{ST + CS + HS\} \tag{1}$$

The framework of the SCSHG-ES is presented, which consists of three stages as illustrated in Figure 1. Its main parts are the proposed E&ISA algorithm and SCSHG model in the first two stages. The following is a detailed introduction to the SCSHG-ES method.

Stage 1: conceptual semantic extending

The first stage is the introduction of the conceptualization mechanism in the offline and online phases. The user's search and the short texts are mapped to unified concept space based

Table 1 Definition of notations

Notation	Definition
S	Given the search by the user, similar to short text in microblog, $S = \{s_1, \ldots, s_n\}$
ST	The plaint text and hashtag extracted from microblog, $= \{ST_1, \ldots, ST_i\}$
CS	The feature of conceptual semantics which generated by Wikipedia, $CS = \{c_1, \ldots, c_i\}$. Divided into three types: CS^S, CS^T, CS^H
HS	The feature of social semantics in hashtags which generated by SCSHG, $HS = \{h_1, \ldots, h_i\}$
CS^S	The set of conceptual semantic feature for the search S
CS^T	The set of conceptual semantics generated for plain text, $CS^T = \{CS^T_1, \ldots, CS^T_n\}$
CS^H	The set of conceptual semantics generated for hashtag, $CS^H = \{CS^H_1, \ldots, CS^H_n\}$
S`	The expanded of S, including S, CS^S and HS
ST`	The expanded of ST, including ST, CS^T and HS

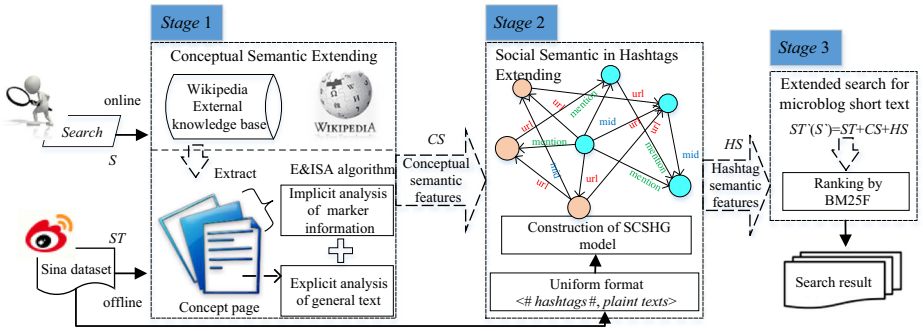


Figure 1 The framework of SCSHG-ES method

on Wikipedia knowledge base, and the corresponding conceptual semantic feature are generated. We propose a new conceptual semantic analysis algorithm E&ISA, which divides the concept page extracted from Wikipedia into marker information and general text. And implicit analysis and explicit analysis are carried out respectively. Then the E&ISA generates a list of the related concepts, represented as a weighted concept vector by sorting according to the relevance of the input. It serves as a conceptual semantic features (*CS*) to extend the original text, denoted to $\{ST + CS\}$.

Stage 2: social semantics in hashtags extending

After the first stage is completed, all the data are represented by corresponding concepts. Furthermore, we extend the *ST* to generate associated hashtag features based on the conceptual semantic features in stage 1 as the social semantics in hashtags. It consists of two parts, an offline phase and an online phase.

Offline phase: Firstly, the data are processed to form a unified format. The most representative top-*n* concepts are selected as the hashtags in the concept set of the plaint text mapping, and the common structure of microblog data are denoted <#*hashtags*#, *plaint texts*>. Then, we construct the SCSHG model through the social connection rules based on co-occurrence of social attributes (such as hashtag, mention and URL). Under the connection of social relations and the conceptual semantic features of hashtags, we can find out the social semantics in hashtags (abbreviated hashtag semantics, *HS*) which is social compact and conceptual semantics similarity. Furthermore, we extend the semantic expression of the short text, which is denoted as $\{ST + HS\}$.

Online phase: Given the search *S*, it is also extended three fields which are similar to *ST*` unifying the structure of the search document. We denote the extended text structure as $S` = \{S + CS + HS\}$. And the conceptual semantic features *CS* is generated in the *stage* 1. So in this phase we need to generate the *HS* for *S*. Since the search has no hashtags, it is similar to the plaint text of microblog. We analyze the similarity of conceptual semantics between a search and the plaint texts which are extracted from microblog. If they overlap more concepts, they have the more similar semantics. We can use the hashtags corresponding to the plaint texts as the associated hashtags of a search, and the hashtags whcih the plaint texts are most similar to the given search are utilized as the social semantics in hashtags.

Stage 3: extended search for microblog short text

Through the above two stages, the representation $S`$ and $ST`$ of the extended features are generated. According to the structured nature of the extended short text, the BM25F [28, 30] ranking algorithm is chosen as the implementation of the extended search. It associates different weights for each field of the extended short text, which reflects the different importance in the search process. Finally, we obtain the search results with highest similarity to the search in the extended semantic space. Following this intuition, we measure the similarity between the $S`$ and $ST`$:

$$Rank\left(S`, ST`\right) = \sum_{t_i \in S`} idf\left(t_i\right).\frac{weight\left(t_i, ST`\right)}{k_1 + weight\left(t_i, ST`\right)} \tag{2}$$

where $idf(t_i)$ is a typical inverse document frequency, and *weight* is the cumulative weight of terms in all fields of extended short text. In the 3.4 section, the Eq. (2) is explained in detail.

3.2 The extension of conceptual semantic features

The extension of conceptual semantic features aims to learn a conceptual representation from raw data. In this section we introduce a combination of explicit and implicit semantic analysis algorithm (E&ISA). The method maps short texts and searches to the concept space based on Wikipedia, and performs matching in that space. Then we can obtain a conceptual semantic features *CS* for *ST* and *S*.

Considering that microblog is a rapidly dissemination social network of emerging information, we select Wikipedia as the knowledge base to help complete the semantic analysis of short text. Because it is a dynamic and fast-growing newsworthy resources, and it is suitable as external reference resources for the search in microblog.

3.2.1 Conceptualization

We preprocess the microblog data and divide them into hashtags and plaint texts. ST_i consists of some terms, which are denoted by $ST_i = \{t_1, \ldots, t_n\}$. We first conceptualize each term to obtain a set of related concepts which are ranked by their representativeness (i.e. $P(c|t)$). The concept mapping probability of terms is $P(c_i|t_i)$, which indicates the representation of a concept. It is given by Eq. (3).

$$P(c|\mathrm{t}) = \frac{count(\mathrm{t}, c)}{\sum_{t \in c_i} count(\mathrm{t}, c_i)} \tag{3}$$

where $count(t, c)$ is the frequency of term t appearing in concept c. c_i belongs to the concept set that t maps in the inverted index.

In the traditional search methods based on Wikipedia, explicit concepts are analyzed in the text of concept page, such as ESA [10]. The concept of title and content are treated equally, represented by spatial vector model. It does not consider the influence of structure and attribute in concept page on conceptual representation. In addition to the page information, concept page of Wikipedia also includes implicit information, for example anchor files and related

events links. Those can help understand the semantics of a given text. Based on the concept of explicit and implicit knowledge of Wikipedia, we present E&ISA algorithm to analyze the explicit and implicit semantics. It realizes the concept mapping of the microblog short text, and transforms text from word vector space to concept space.

In order to construct a conceptual semantic space model which maps the ST into a list of weighted relevance concepts, we first divide the Wikipedia page into two parts to conduct concept mapping respectively. Then we obtain the concepts combined with the two parts of conceptual score. The ST is vectorized as $ST_i = \{w_{c1}, w_{c2}, \ldots, w_{ck}\}$. Each item is the corresponding weight of the ST under the concept c_i, which represents the correlation strength between the concepts and terms, and it is estimated as:

$$w_{ci} = \log \sum_{t_i \in ST_i} T_i \cdot idf(t_i) \cdot icf(c_i) \tag{4}$$

T_i means the representative score of the t_i in short text ST_i corresponding to the concept c_i, and $T_i = \{p(c_i|t_i, ST_i)\}$. The $idf(t_i)$ is the inverse document frequency of t_i. The $icf(c_i)$ is the inverse concept frequency of c_i. It is calculated similar to idf, which indicates the representation of the concept c_i.

3.2.2 The conceptual semantic features generation

During the implementation of E&ISA algorithm, we extract the titles, texts, anchor files and internal links. They are divided into explicit information part (i.e. general text domain, including the texts) and implicit information part (i.e. marking information domain, including the titles, anchor files and internal links). The contribution of these two parts to the concepts of the article is different. The implicit information contains some related concepts of descripted in link pages, enhancing the ability of the conceptual representation of the page. Therefore, the implicit information is more representative than the explicit information in the representation of the concept.

In order to show the importance of implicit information, we give the terms in the field of implicit information a higher weight than the normal terms in the explicit information in conceptual mapping. Firstly, we create inverted indexes in explicit and implicit concept sets by using the Apache Lucene[1] to accelerate the conceptual mapping of short text respectively. Then, the weights of conceptualized vectors are returned by retrieving the terms. In this operation, both of explicit and implicit information are effectively incorporated into the calculation of the semantic similarity. If the terms belong to the general text domain of concept c_i, they will be analyzed and calculated as the explicit information. If the terms appear in the marking part of the concept c_i, they will be treated as the implicit information. So they will have different conceptual importance corresponding to different positions. The λ is introduced to control the influence of the two parts on the representative between the terms and the concepts, and the weight of the concept c is obtained as:

$$T_{\mathrm{i}} = p(\mathrm{c}_i|\mathrm{t}_i) = \lambda \cdot p^e(\mathrm{c}_i|\mathrm{t}_i) + (1-\lambda) \cdot p^i(\mathrm{c}_i|\mathrm{t}_i) \tag{5}$$

where $p^e(c_i|t_i)$ and $p^i(c_i|t_i)$ denote the representation probability of concept c_i in the explicit and implicit information corresponding to t_i respectively. The λ is an element parameter, and

[1] http://lucene.apache.org

$\lambda \in [0,1]$. It reveals the importance of assigning weights to explicit and implicit information, and we set it to 0.7.

Algorithm 1 E&ISA

Input:

The given search S and text data of microblog $\mathcal{D}$, commonly denoted as *text* during this operation

The parameters λ of explicit and implicit weight distribution

The number of concepts generated for each text K

Output:

The concept set CS^S, CS^H and CS^T

1: Extract Wikipedia concept page generated concept set $C = \{C_1, \cdots, C_M\}$

2: Separate each concept page from explicit and implicit information and obtain $C^{ex} = \{C_i^{ex}\}_{i=1}^{M}$ and $C^{im} = \{C_i^{im}\}_{i=1}^{M}$, respectively.

3: $\{c_i\}_{i=1}^{K} = \varnothing$;

4: **for** each *text* **do**

5: $InvIndex[\, C_i^{ex} \,] \leftarrow createInvIndex\,(text, C^{ex})$, $InvIndex[\, C_i^{im} \,] \leftarrow createInvIndex\,(text, C^{im})$;

6: **for** $t_i \in text$ **do**

7: $c^{ex} = \{c_1, \ldots, c_n\} \leftarrow retrieval(t_i, C^{ex})$, $c^{im} = \{c_1, \ldots, c_m\} \leftarrow retrieval(t_i, C^{im})$;

8: **if** $c_i \in c^{ex} \cap c^{im}$ **then**

9: $C_i = \{c_1, \ldots, c_i\} \leftarrow combine(t_i, C)$;

10: **if** $c_i \in \underset{1 \le i \le n}{\cap} C_i$ **then**

11: $\{c_i\}_{i=1}^{K} \leftarrow combine(text, C)$;

12: **return** $\{c_i\}_{i=1}^{K}$ (corresponding to the CS^S, CS^H and CS^T respectively)

After the short text is conceptualized, it is converted into the conceptual space to extend the conceptual semantic features of the short text. In this space, a unified extended form is obtained $\{ST + CS\}$. For example, considering short text "The big explosion in Tianjin", the generated *CS* is "explosion accident of dangerous chemical warehouse in Tianjin port in 2015", "serious fire and explosion accident in Tianjin port", "the mushroom cloud", "Tianjin", etc. Finally, in order to make the algorithm more formal we summarize the E&ISA in Algorithm 1.

3.3 The extension of hashtag semantic features

After we get conceptual semantics through the first phase, this section further extends the short text using social semantics in hashtags. In the concept space, the different concepts are independent of each other. By constructing a SCSHG model, the concepts are connected through social relations, including hashtag and other auxiliary attributes to form a graph structure. It maintains the topic consistency. In the short text of microblog, the "mention" refers to an active microblog user who discusses the same topic or has a certain authority on the issue. It can be used as a link between microblog entities (messages, i.e. short texts), so as

to the messages can be linked by user's topics of common interest. Similarly, the messages containing the same "URL" discuss and focus on the same link content. Through this information, we can form a potential association for the entire microblog network to represent and calculate the natural semantics of social networks in the form of graphics. Taking into account the associations constructed by social features, the correlation among hashtags is organized in a conceptual space.

On the basis of hashtag graph, the concept space is introduced. In the generation of associated hashtags, we use shared concepts and the social connection rules to achieve the connection among nodes. Then the set of semantic consistency hashtags is generated to further enhance the semantics of short text.

3.3.1 The SCSHG model construction

After short texts and hashtags are transformed into conceptual semantics, they are mapped into a unified concept space. Because the microblog short text with hashtags only accounts for a small part, the top-*n* related concepts will be selected as the hashtags for plaint text (the short texts of microblog without hashtags). The microblog data space is represented as a unified format {*#hashtags#*, *plaint texts*}. In this way, we construct a complete and compact graph $G = <V, E, w>$, where V is the set of nodes represented by hashtags, and $E = \{(v^i, v^j) | v^i, v^j \in V\}$ is the set of edges. For each edge $(u, v) \in E$, $w(u, v)$ is the semantic similarity between u and v, which ranges over (0, 1). The edge denotes social relationship between hashtags. To combine conceptual semantics with social relations in microblog, the social connection rules between hashtags are defined as follows:

> **Rule 1** *Mid co-occurrence*. If the hashtags appear in the same short text, there is a connection between hashtag nodes. It indicates that the two hashtags have the same topic tendency.
> **Rule 2** *Mention co-occurrence*. If the hashtags exist in the short texts of the same person or organization mentioned, we establish a connection between the hashtag nodes. It shows that the contents of two hashtags are related to the same person mentioned, the hashtags are similar in semantic in this case. For example, if the "Tianjin Fire" is mentioned at the same time in two microblog contents, they both involve the topic "explosion in Tianjin".
> **Rule 3** *URL co-occurrence*. If the hashtags appear in the short text with the same URL, then the hashtag nodes form a connection edge. This is because the text that contains the same link may discuss the same event or topic.

Importing SCSHG into Neoj4 graph database [21] which is suitable for social network graph structure, we construct the SCSHG model. It makes the nodes more intuitive and flexible for association operations. Further, we measure the associated hashtags in the concept space and generate the associated hashtag features of the short text. It is extended as $\{ST + HS\}$. The w is the set of edge weights, which reflects the correlation between the connecting hashtags. And it is related to the average cumulative weight of each edge on a variety of social relationships and the degree of overlap concepts of hashtags (i.e. CS^H). The more the overlap of concepts between the hashtags, the more relevant the semantics of the hashtags.

Figure 2 is the SCSHG model with some nodes in the same event shown enlarged. The nodes are divided into two types of hashtag and concept (where the hashtag nodes are

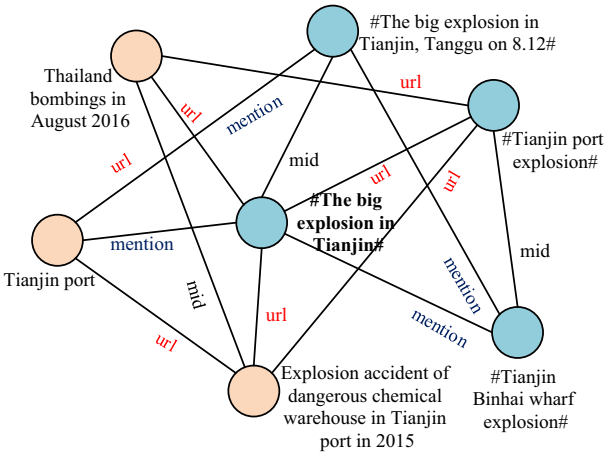


Figure 2 Instance representation of partial nodes enlarged in the same event of SCSHG model

indicated by # #). Each node contains attributes *mid*, *mention*, *URL* and *concept*. Such as #The big explosion in Tianjin # is a hashtag node, its four attributes are as follows:

mid = (CvA0R2lpe, CvA1MFlr1, CvA22mx7H, CvA2BwqdL, CvA2XCYkS, etc.)
mention = (Tianjin fire protection, Sina Tianjin, China news net, people's daily)
URL = (http://t.cn/RL3Sxmg, http://t.cn/ RL3J6I4)
concept = (Explosion accident of dangerous chemical warehouse in Tianjin port in 2015, Explosion accident, Tianjin, Tianjin refueling, Mushroom cloud, etc.)

Since it matches Rule 1 with the hashtag node "#The big explosion in Tianjin, Tanggu on 8.12#", the two nodes are connected. It also connects to other hashtag nodes including #Tianjin port explosion# and #Tianjin Binhai wharf explosion# because of the different social connection rules. In addition, it meets Rule 2 with the concept node "Tianjin Port" and there is an edge between them. The connected concept nodes also include "Thailand bombings in August 2016" and "Explosion accident of dangerous chemical warehouse in Tianjin port in 2015".

3.3.2 The associated hashtag feature generation

In this section, we need to generate the associated hashtags of a search and plaint text, which are done online and offline phase respectively. They are based on conceptual semantic features, and the feature of the search and plaint texts are represented as CS^S and CS^T.

First, we generate the hashtag semantic feature of the search, and accomplish the extension of the *HS*. Evaluating semantic similarity between CS^S and C^T, we select the hashtags corresponding to the plaint texts with CS^T having high similarity with CS^S. The semantic similarity is defined as Eq. (6) according to the common concepts in CS^S and CS^T:

$$Sim\left(CS^S, CS_i^T\right) = \frac{\sum_{c_i \in CS^S \cap CS_i^T} p\left(c_i \middle| CS^S\right) \cdot p\left(c_i \middle| CS_i^T\right)}{\left|CS^S \cap CS_i^T\right|} \tag{6}$$

Table 2 The instance of extended associated hashtag features generated by hashtag or plaint text

Hashtag/plaint text	Association hashtags feature
#The big explosion in Tianjin, Tanggu#	#Explosion accident of dangerous chemical warehouse in Tianjin port in 2015#, #Tianjin Binhai wharf explosion#, #Explosion accident in Tianjin port#, #Tianjin explosion#, #Tanggu explosion #, etc.
Rainstorm	Concepts "Floods in southern China in 2016", "heavy rainfall", #Rainstorm in Hubei#, #Wuhan rainstorm#, #Rainstorm direct seeding#, #Rainstorm disaster#, etc.

$|CS^S \cap CS_i^T|$ is the number of common concepts in both sets, $p(c_i|CS^S)$ and $p(c_i|CS^T)$ denote the representative score of common concept c^i in the corresponding set.

Second, the SCSHG model is used to calculate the associated hashtags for the plaint texts, and we define semantic coherence scores to evaluate the correlation probability between hashtag n and target hashtag o. In the SCSHG, there is close social semantic similarity between adjacent nodes, thus we can match a set of semantic consistency hashtags for each target node. The semantic consistency scores of target hashtag o and its neighbor hashtag n are computed as:

$$S(n_i|o) = e(n_i|o) + Sim(n_i|o) \tag{7}$$

where $e(n_i,o)$ is the weight of the edge between n_i and o. It is the average of the weights given by the three connection rules. For each rule, the weight of the edge is the ratio of the number of conditions in which two nodes co-occurrence to the number of the target node that contains the attribute, i.e. it is proportional to the number of co-occurrence *mids* in both hashtag nodes and *mids* contained in target nodes.

The semantic similarity measure of two hashtags is the same as the calculation between a search and plaint text, i.e. Eq. (4). It compares the conceptual semantics between nodes generated by the E&ISA algorithm. The conceptual semantics is CS^H or CS^T (when the node is a concept). In other words, the number of overlapping concepts of both hashtags can be used to evaluate their semantic similarity. The more concepts two nodes overlap, the more similar semantic description. Then top-k hashtags with high semantic consistency scores are selected as hashtag semantics *HS*. Table 2 is an instance of associated hashtag features generated by a hashtag or plaint text in SCSHG model. The *HS* generated by plaint text contains two concepts.

3.4 Extended search of microblog short texts

On the basis of BM25F method [29], this paper takes into account the role of each field in the structured text which is different on the search, and gives it different semantic information. The hashtags play an important guiding and aggregation role in the consistency of microblog event development. Since conceptual semantics help to understand texts, we give it a higher weight than the original text. The cumulative weights of the terms in all fields are obtained as:

$$weight(t_i, ST) = \sum_{c \in ST} \frac{occurs_{t_i,c}^{ST}.\mathrm{boost}_c}{\left((1-\mathrm{b}_c) + \mathrm{b}_c.\dfrac{l_c}{avl_c}\right)} \tag{8}$$

where l_c is the length of the field, avl_c is the average length of the c, and $occurs_{t_i,c}^{ST}$ is the number of t_i appearing in the field c for the short text ST. The b_c is regulatory factor related to field length. Followed by [28], we set $b_{HS} = 0.4$, $b_{CS} = 0.4$ and $b_{ST} = 0.3$. The $boost_c$ is the enhancement factor used on the field c, which is set to $HS = 5$, $CS = 3$ and $ST = 2$ respectively. In the nonlinear saturation function $weight/(weight + k1)$, $k1 = 1.7$. The ranking function is given as Eq. (2), where the calculation of $idf(t)$ is shown in Eq. (9):

$$idf(t_i) = \log \frac{N - df(t_i) + 0.5}{df(t_i) + 0.5} \tag{9}$$

N is the number of documents in the dataset, and df is the number of documents with t appears.

4 Experimental and analysis

In this section, we show experiments on three microblog datasets to evaluate the performance of our method in several aspects. We first conduct analysis on the influence of different parameter settings on the SCSHG-ES method, and then investigate the performance of SCSHG-ES by comparing with other extended search methods. Meanwhile, we provide the results with discussions.

4.1 Datasets

4.1.1 Wikipedia

This paper uses the Wikipedia data released on July 01, 2017 version at "https://dumps.wikimedia.org/zhwiki" with Wikidump format.[2] They are stored in the database respectively, such as page.sql, interlinks.sql. We extract the required information through multi-table operation, including title, content, anchor file and internal links. The concept page is divided into the domain of text description (page_id, content) and the domain of marking information (title, anchor_text, interlinks). In order to complete the mapping and matching between terms and concepts, we build an inverted index of microblog data in both domains using the Apache Lucene respectively. The preprocessing is carried out as follows:

Firstly, the concepts filtering. The downloaded data contain non-concept pages (such as Talk) that have no effect on semantic interpretation of text. Short pages are difficult to achieve semantic extension and isolated pages have too few links in the relational structure. To reduce the noise and interference to the conceptual operation, we delete the non-concept pages and the pages with less than 200 words or 3 links.

Secondly, conceptual information extraction. The candidate concept pages are parsed by wikiExtractor[3] as the text format, and the texts are converted into simplified chinese by OpenCC. The size of the downloaded data is 1.4G. After parsing, there are 948,835 articles and there are 167,328 candidate concepts remaining.

[2] https://en.wikipedia.org/wiki/Wikipedia:Database_download
[3] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

4.1.2 Datasets of microblog and preprocessing

The national security related data in Sina microblog were crawled through relevant keywords. The datasets containing precise events and small noise are created by search in the form of keywords and their combinations. We crawled 168,199 data from two major events and processed them into the following three datasets. Single event dataset 1 is obtained from the keywords "rainstorm, Hubei, Wuhan, flood control, disaster" and their random combination; Single event dataset 2 is obtained by searching for the keywords "Tianjin, Tanggu, explosion, warehouse, Binhai New Area" and their random combination; Mixed event dataset 3 is a merge of dataset 1 and dataset 2.

For each dataset, microblog content is extracted to obtain social auxiliary information, including plaint texts, hashtags, mentions (@) and URLs, which can be empty. The plaint texts and hashtags are converted into simplified chinese and achieved word segmentation. In conceptual mapping, only the keywords that describe events are conceptualized to avoid the impact of generating noise concepts in search. The amount of social auxiliary information contained in the extracted microblog content is shown in Table 3. It is imported into the graph database Neoj4 to form the nodes and attributes of connection graph.

4.2 Experimental setting

4.2.1 Search settings and baseline methods

We invite experienced microblog users to participate in search tasks, and evaluate the results of experiment. Each person gives any ten searches related to the event as a search collection. We evaluate the validity of SCSHG-ES by comparing it to the baseline methods. Each method returns top-K search results to form a result collection. In order to avoid a cognitive bias among the learning effects caused by the search, the result collection is anonymous and randomly combined and assigned to the experimental participants. Then, the K most related with the search results are specified by participants. Finally, we calculate the correlation between the results marked by the user and the K results returned by each method. The baseline methods are described as follows:

ESA-ES [8]: It is the concept-based extended search method which adopts the ESA to generate the concepts of text through explicit semantic analysis, we denote the methods as ESA-ES.

E&ISA-ES: By the improvement of ESA algorithm, we propose the semantic analysis method E&ISA which combines explicit and implicit information. It is used for concept-based extended search. We denote this method as E&ISA-ES. This is the first stage of the SCSHG-ES method, because it has no associated hashtags.

Table 3 The social auxiliary information statistics in microblog datasets

Dataset	#weibo	#hashtag	#mention@	#URL
Dataset 1	81,270	13,156(16%)	7398(9.1%)	11,031(13.5%)
Dataset 2	86,929	19,124(22%)	8997(10.3%)	13,039(15%)
Dataset 3	168,199	32,280(19%)	16,395(9.7%)	24,070(14%)

Topic-ES [43]: The topic-based extended search. It uses topic words instead of concepts. The LDA [6] generates topic distributions to represent the short texts. It is denoted as Topic-ES.

SEMD-ES [3]: It is an extended search based on semantically enriched microblog document. Its document contains plain text, hashtag, segmentation hashtag, lead paragraph of Wikipedia linked entities in segmented hashtags and texts. We denote the method as SEMD-ES.

4.2.2 Evaluation metric

The evaluation metric of information retrieval mainly includes the ability to search related documents and sort the relevant text correctly [13]. We have a comprehensive evaluation on the performance of the search method by using the precision (P@K), the mean average precision (MAP) and the normalized discount cumulative gain (NDCG), where K is the threshold of the search result.

The precision is the proportion of the relevant documents in the return result, which is defined as:

$$\text{precision} = \frac{tp}{tp + fp} \tag{10}$$

where tp is the number of correct document (i.e. positive class related to a search) in the return results, fp is the number of incorrectly document assigned to the positive class. The sum of t_p and f_p is the total number of results list.

MAP takes into account the location factor on the basis of precision. The mean of average precision for the search set is determined by the average of the MAP values from all individual search. For a single search, the calculation method of the average precision is:

$$MAP = \frac{1}{r}\sum_{i=1}^{r} \frac{i}{\text{the location of i}^{th}\text{ relevant document}} \tag{11}$$

NDCG is the index of continuous value, which is calculated based on the top-K retrieval results as Eq. (12):

$$NDCG(\mathrm{Q},\mathrm{k}) = \frac{1}{|\mathrm{Q}|}\sum_{j=1}^{|\mathrm{Q}|} Z_{j,k} \sum_{m=1}^{k} \frac{2^{R(\mathrm{j},\mathrm{d})}-1}{\log(1+\mathrm{m})} \tag{12}$$

where $R(j, d)$ is the correlation score of the document d given by the evaluator to the search j (usually a probability), m is the location where the document is returned.

4.3 The parameter setting and discussion

The extended search methods will achieve their best performance when the number of extended words is set to 20 [5, 27]. Therefore, we set the number of extension words k to 20 in concept-based extension of ESA-ES and E&ISA-ES methods. In the Topic-ES method, the parameters of topic model LDA is $\alpha = 50/l$ where we set the topic number $l = 10$, $\beta = 0.01$, the iterations of Gibbs sampling are set to 1000. Each subject returns 20 topic words for

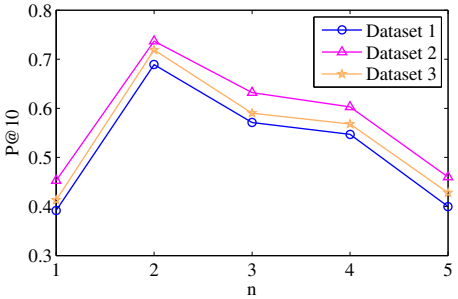


Figure 3 The influence of parameter n on the precision of SCSHG-ES method

extension search. In addition, the number of words used to extend after segmentation is not consistent because of the variable length of concepts and hashtags in the SCSHG-ES method. Therefore, we set the empirical parameter $k = 5$ for the number of the conceptual features and the associated hashtag feature extensions.

4.3.1 The influence of n on SCSHG-ES method

In Section 3.3, the parameter n will affect the compactness of hashtag graph constructed in the SCSHG-ES method, which is the number of concepts chosen as the hashtags of plaint text. It also plays a role in the generation of associated hashtag features. Furthermore, the search efficiency of SCSHG-ES method is affected by the change of n to a certain extent. For the selection of parameter n, we experiment on three datasets and get the optimal value. Figure 3 shows the sensitivity of the SCSHG-ES method for the search performance P@10 under the different values of the n.

In Figure 3, the SCSHG-ES method performs better on three datasets when $n = 2$. It shows that the concepts are a kind of semantic summarization of short text, similar to the role of hashtag in microblog. Two concepts are used as hashtags for each short text, which can fit the actual number of hashtag in microblog environment. In this case, social networks are tightly connected and have no redundant hashtags, so we set n to 2 in the follow-up experiments.

In addition, our method performs the best performance on dataset 2, and the search result in dataset 3 is superior to dataset 1. As can be seen from the number of social auxiliary information in each dataset in Table 3, the search performance of the SCSHG-ES method is positively related to the proportion of the social auxiliary information contained in the microblog datasets. This also shows that the number of auxiliary information plays a certain

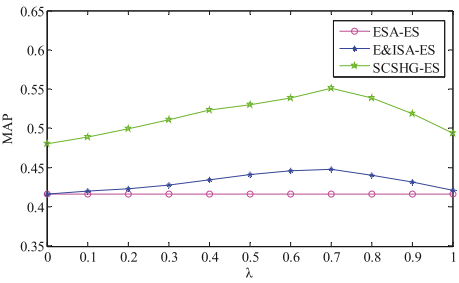


Figure 4 The influence of λ on the MAP of SCSHG-ES and correlation contrast methods

role in the performance of extended search for SCSHG-ES method. And the larger the proportion, the better the performance of guiding search.

4.3.2 The influence of λ on the SCSHG-ES and correlation contrast methods

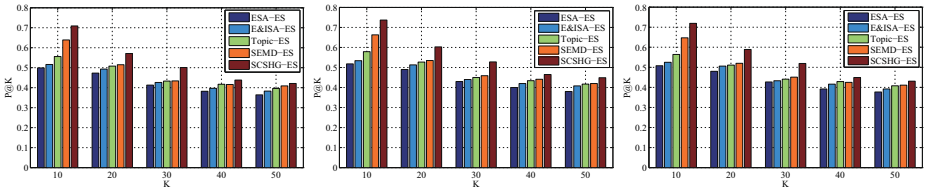
The search precision of the proposed method will change with the distribution of the explicit and implicit weight parameter λ in the extension of the conceptual semantic features (as the Eq. 6). In this experiment, we study how the search performance is affected by the λ. The ESA-ES, E&ISA-ES and SCSHG-ES methods are carried out on dataset 2. The reason we choose dataset 2 is that it has the largest percentage of social auxiliary information. This allows our method to maximum advantage. Under it best conditions, we can evaluate the influence of parameter λ and reduce the interference of other factors on the experiment. Figure 4 shows the sensitivity results of SCSHG-ES and correlation contrast methods to the MAP in different parameter λ. We set the number of returned results K to 10, and analyze the effect of weight distribution on the search.

The MAP value of the ESA-ES method is not affected by λ in Figure 4, which illustrates that the method conducts only explicit analysis on the text. If the implicit information is ignored (when $\lambda = 0$), then E&ISA-ES is equivalent to ESA-ES, they will have the same search performance. In contrast, the MAP of E&ISA-ES and SCSHG-ES methods significantly changes with the λ and they achieve the optimal as $\lambda = 0.7$. It explains that the effectiveness of the improvement of conceptual analysis is obvious by adding the parameter λ. In other words, the implicit analysis in the conceptual mechanism is effective. Moreover, the MAP of E&ISA-ES and SCSHG-ES increases with parameter λ when $\lambda < 0.7$, otherwise it will decrease. So it reaches the optimal value at 0.7. Because the extension of associated hashtags is included in the SCSHG-ES method, it is slightly better than baseline methods. So the introduction of λ is meaningful and it can enhance the effectiveness of our method. We can get the conclusion that the extension of the social semantics in hashtags in this paper can further improve the search result of a simple concept-based extension method.

In order to more clearly demonstrate the effect of introducing the parameter λ, we generate the top-5 conceptual features for terms or short texts with ESA and E&ISA as shown in the Table 4. It can be seen from that E&ISA algorithm can dig out more event linkage concepts, such as "Floods in southern China in 2016" and "Explosion accident of dangerous chemical warehouse in Tianjin port in 2015". It is closer to the semantics of terms or short texts expression than ESA. However, the ESA algorithm only performs explicit semantic analysis, the conceptualization results tend to be more literal.

Table 4 Comparison the extended conceptual features between ESA and E&ISA

Term/short text	Concepts from ESA	Concepts from E&ISA
Rainstorm	Rain, Wind hail, Flood, Storm, mud-rock flow	Floods in southern China in 2016, Rain, Flood, Mud-rock flow, Heavy rainfall
The big explosion in Tianjin, Tanggu	Explosion accident of dangerous chemical warehouse in Tianjin port in 2015, Explosive equivalent, Tianjin port, Tianjin	Explosion accident of dangerous chemical warehouse in Tianjin port in 2015, serious fire and explosion accident in Tianjin port on 8.12, The mushroom cloud, Tianjin



(a) Dataset 1 (b) Dataset 2 (c) Dataset 3

Figure 5 Comparison of P@K between SCSHG-ES and baseline methods on three datasets

4.4 Performance evaluation

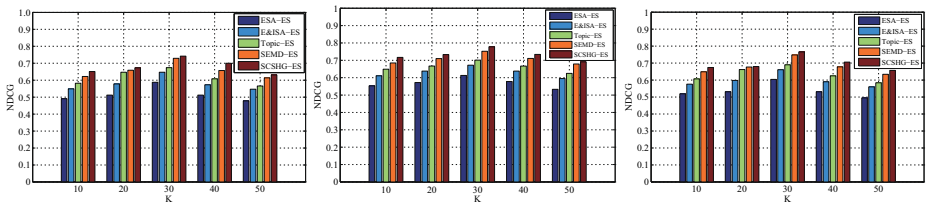
In this section, we describe a series of experiments that evaluate the performance of the proposed extended search, and verify the performance of SCSHG-ES method on P@K, MAP and NDCG. By comparing with the baseline methods, the validity of our method is further analyzed.

4.4.1 Comparison on the P@K

We further verify the search performance of SCSHG-ES method, and take the parameter λ as 0.7, which makes the SCSHG-ES method reach the best condition and compare with the baseline methods on different datasets. Figure 5 shows the comparison of P@K metric, the precision of each method reaches the peak value when the number of returned results $K = 10$. With the increase of K, the precision of all methods decreases gradually. Specifically, the SCSHG-ES method performed best. SEMD-ES and Topic-ES methods have little difference in search precision, which are better than E&ISA-ES. The worst is ESA-ES method. It shows that the simple concept-based method is not ideal, and the method of using hashtags has better precision.

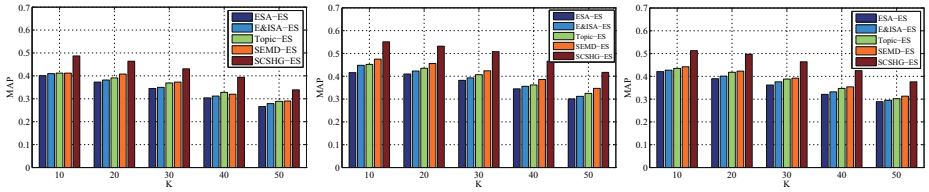
4.4.2 Comparison on the NDCG

Figure 6 is the comparison of NDCG between SCSHG-ES and baseline methods in three datasets, where λ is set to 0.7. Experimental results show that all methods achieve the best performance when the number of returned results K is 30. Figure 6b is the results on the dataset 2, the SEMD-ES and SCSHG-ES methods have better NDCG than the other datasets. It illustrates that the amount of social auxiliary information contained in dataset affects the search results for both methods. Among them, the SCSHG-ES performance changes more than the other methods on three datasets, which shows that this method is effective in extended search by using microblog auxiliary information.



(a) Dataset 1 (b) Dataset 2 (c) Dataset 3

Figure 6 Comparison of NDCG between SCSHG-ES and baseline methods on three datasets



(a) Dataset 1 (b) Dataset 2 (c) Dataset 3

Figure 7 Comparison of MAP between SCSHG-ES and baseline methods on three datasets

4.4.3 Comparison on the MAP

Figure 7 is the experimental result of the SCSHG-ES and the baseline methods under the three datasets at the MAP value, we set the λ to 0.7. All methods have the best MAP value when the number of returned results $K = 10$, and the MAP value of SCSHG-ES method in the three datasets is significantly better than the baseline methods. The MAP value of SCSHG-ES and SEMD-ES on dataset 2 is better than that of the other datasets. This phenomenon shows that the hashtag information is used in those methods, and the hashtag proportion will affect the performance of search.

For the comprehensive analysis from Figures 5, 6 and 7, our method is better than the baseline methods on all datasets for P@K, NDCG and MAP, followed by the SEMD-ES method using the hashtag information. The next is the Topic-ES method. The E&ISA-ES is better than ESA-ES, which shows that the effectiveness of the explicit and implicit information is dealt with separately. The change of dataset has the most obvious influence on SCSHG-ES and SEMD-ES methods, because these methods use hashtags and other auxiliary information. The overall performance of the method is affected by the percentage of the hashtags in dataset.

Table 5 displays the average value of the performance of the search experiment conducted on three datasets. It can be seen from Table 5 that the average search performance of SCSHG-ES method is the best, which shows that the method has good stability. The performance of E&ISA-ES is better than the ESA-ES. The extended search methods which only use the concept (ESA, E&ISA-ES) or topic (Topic-ES) are not as good as the algorithm with hashtags (SEMD-ES, SCSHG-ES) with search accuracy. SCSHG-ES method complements the hashtags and adds the social semantic extension, thus the meaning of short text is deeply understood. Furthermore, it improves the semantic search of the short text effectively. However, the SEMD-ES does not deal with a large number of the short texts without hashtags, and does not take into account the semantic information of each field and the social relationships. So the effect is not as good as SCSHG-ES.

5 Conclusion

In this paper, we propose a new extended search method SCSHG-ES for microblog short text, aiming to find more semantic information for similarity inference in search. It seeks help in semantic coherence between the concepts and topic guidance of the hashtags. We first perform the semantic analysis by E&ISA algorithm which combines the explicit and implicit concepts in Wikipedia, and generate the conceptual semantic features. Then the SCSHG model is proposed to realize the extension of social semantics in hashtags. Finally, the experiments

Table 5 Comparison of the average search performance between SCSHG-ES and baseline methods

	P@K					NDCG@K					MAP@K				
	10	20	30	40	50	10	20	30	40	50	10	20	30	40	50
ESA-ES	0.508	0.481	0.423	0.391	0.373	0.514	0.531	0.601	0.533	0.495	0.412	0.391	0.363	0.323	0.285
E&ISA-ES	0.525	0.504	0.433	0.411	0.394	0.572	0.598	0.66	0.593	0.560	0.428	0.402	0.373	0.333	0.295
Topic-ES	0.566	0.515	0.441	0.426	0.407	0.606	0.652	0.688	0.626	0.585	0.433	0.414	0.388	0.345	0.305
SEMD-ES	0.649	0.523	0.448	0.427	0.413	0.645	0.676	0.743	0.675	0.635	0.443	0.429	0.396	0.353	0.316
SCSHG-ES	0.721	0.587	0.516	0.450	0.433	0.674	0.688	0.762	0.706	0.653	0.517	0.497	0.467	0.428	0.377

on microblog datasets demonstrate that the SCSHG-ES can fully understand the semantics of short text and effectively improve the quality of search. In the future, we are intended to apply more effective social attribute characteristics to text and image searching. Thus, the concept and social semantics will be applicable to a wider range of search tasks.

Acknowledgments This work was supported by the National Natural Science Foundation of China (NSFC) under Grant (No.61320106006, No.61532006, No.61772083, No. 61502042).

References

1. Albishre, K., Li, Y.F., Xu, Y.: Effective pseudo-relevance for microblog retrieval. In: Australasian Computer Science Week Multi conference, pp.51. ACM (2017)
2. Bandyopadhyay, A., Ghosh, K., Majumder, P., Mitra, M.: Query expansion for microblog retrieval. Int. J. Web Sci. **1**(4), 368–380 (2012)
3. Bansal, P., Jain, S., Varma, V.: Towards Semantic Retrieval of Hashtags in Microblogs. In: Proceedings of the 24th International Conference on World Wide Web, pp.7–8. ACM (2015)
4. Cao, G., Nie, J.Y., Gao, J.F., Stephen, R.: Selecting good expansion terms for pseudo-relevance feedback. In: Proceedings of the 31st annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.243–250. ACM (2008)
5. Cindy, H., William, A.W., Malik, M.I.: Query expansion. In: Encyclopedia of Social Network Analysis and Mining, pp. 1455–1455. Springer, New York (2014)
6. David, M.B., Andrew, Y.N., Michael, I.J.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
7. Doroudian, M., Akbari, R.: Using ontologies for developing a search mechanism in social networks. In: International Industrial Engineering Conference, pp.114–120. IJWA (2014)
8. Egozi, O., Markovitch, S., Gabrilovich, E.: Concept-based information retrieval using explicit semantic analysis. ACM Trans. Inf. Syst. **29**(2), 1–34 (2011)
9. Fan, F.F., Qiang, R.W., Lv, C., Yang, J. W.: Improving microblog retrieval with feedback entity model. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp.573–582. ACM (2015)
10. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: international joint conference on Artificial intelligence, pp.1606–1611. IJCAI (2007)
11. Gao, L.L., Wang, Y., Li, D.S., Shao, J.M., Song, J.K.: Real-time social media retrieval with spatial, temporal and social constraints. Neurocomputing. **253**, 77–88 (2017)
12. Guo, D., Gao, P.F.: Complex-query Web image search with concept-based relevance estimation. World Wide Web. **19**(2), 247–264 (2016)
13. Hawking, D., Craswell, N., Bailey, P., Griffihs, K.: Measuring search engine quality. Inf. Retr. **4**(1), 33–59 (2001)
14. Herranz, L., Jiang, S.Q., Li, X.Y, Scene recognition with CNNs: objects, scales and dataset Bias. In: Computer Vision and Pattern Recognition, pp,571–579. CVPR (2016)
15. Hong, K.J., Kim, H.J.: A Semantic Search Technique with wikipedia-based text representation model. In: 2016 International Conference on Big Data and Smart Computing, pp.177–182. IEEE (2016)
16. Hua, W., Wang, Z.Y., Wang, H.X., Zheng, K., Zhou, X.F.: Understand short texts by harvesting and analyzing semantic knowledge. IEEE Trans. Knowl. Data Eng. **29**(3), 499–512 (2017)
17. Huang, G.Y., He, J., Zhang, Y.C., Zhou, W.L., Liu, H., Zhang, P., Ding, Z.M., You, Y., Cao, J.: Mining streams of short text for analysis of world-wide event evolutions. World Wide Web. **18**(5), 1201–1217 (2015)
18. Jia, Y., Gan, L., Li, A.P., Xu, J.: Research progress and development trend of online social network smart search. J. Commun. **36**(12), 9–12 (2015)
19. Jiang, D., Leung, W.T., Ng, W.: Query intent mining with multiple dimensions of Web search data. World Wide Web. **19**(3), 475–497 (2016)
20. Jiang, Y.C., Bai, W., Zhang, X.P., Hu, J.J.: Wikipedia-based information content and semantic similarity computation. Inf. Process. Manag. **53**(1), 248–265 (2016)
21. Kang, J.H., Luo, Z.X.: Research on RDF data storage based on graph database Neo4j. Inf. Technol. **6**, 031 (2015)

22. Kiesel, J., Potthast, M., Hagen, M., Stein, B.: Spatio-temporal analysis of reverted wikipedia edits. In: International AAAI Conference on Web and Social Media, pp,122–131. AAAI (2017)
23. Kotov, A., Zhai, C.X.: Tapping into knowledge base for concept feedback: leveraging conceptnet to improve search results for difficult queries. In: Proceedings of the fifth ACM International Conference on Web Search and Data Mining, pp.403–412. ACM (2012)
24. Lu, Z., Zha, H.Y., Yang, X.K., Lin, W.Y., Zheng, Z.H.: A new algorithm for inferring user search goals with feedback sessions. IEEE Trans. Knowl. Data Eng. **25**(3), 502–513 (2013)
25. Luo, Z.C., Yu, Y., Osborne, M., Wang, T.: Structuring tweets for improving twitter search. J. Assoc. Inf. Sci. Technol. **66**(12), 2522–2539 (2015)
26. Miyanishi, T., Seki, K., Uehara, K.: Time-Aware Latent Concept Expansion for Microblog Search. In: International AAAI Conference on Web and Social Media. AAAI (2014)
27. Ogilvie, P., Voorhees, E., Callan, J.: On the number of terms used in automatic query expansion. Inf. Retr. **12**(6), 666–679 (2009)
28. PérezAgüera, J.R., Arroyo, J., Greenberg, J, Iglesias, J.P., Fresno V: Using BM25F for semantic search. In: Proceedings of the 3rd International Semantic Search Workshop, pp.2. ACM (2010)
29. Péreziglesias, J., Pérezagüera, J.R., Fresno, V., Feinstein, Y.Z.: Integrating the probabilistic models BM25/ BM25F into Lucene. Comput. Sci. **5046**, (2009)
30. Shi, Z.D., Keung, J., Song, Q.B.: An empirical study of BM25 and BM25F based feature location techniques. In: Proceedings of the International Workshop on Innovative Software Development Methodologies and Practices, pp.106–114. ACM (2014)
31. Song, X.H., Jiang, S.Q., Herranz, L.: Multi-scale multi-feature context modeling for scene recognition in the semantic manifold. IEEE Trans. Image Process. **26**(6), 2721–2735 (2017)
32. Sun, J., Xu, J., Zheng, K., Liu, C.: Interactive spatial keyword querying with semantics. In: ACM on Conference on Information and Knowledge Management, pp, 1727–1736. CIKM (2017)
33. Tran, T., Tran, N.K., Asmelash, T.H., Jäschke, R.: Semantic annotation for microblog topics using wikipedia temporal information. EMNLP (2017)
34. Wang, Z., Zhao, K., Wang, H., Meng, X.F., Wen, J.R.: Query understanding through knowledge-based conceptualization. In: International Conference on Artificial Intelligence, pp.3264–3270. IJCAI(2015)
35. Wang, X.Y., Zheng, Y.Q., Xiao, Y.H.: Entity-relation modeling and discovery for smart search. J. Commun. **36**(12), 17–27 (2015)
36. Wang, Y., Liu, J., Huang, Y.L., Feng, X.: Using hashtag graph-based topic model to connect semantically-related words without co-occurrence in microblogs. IEEE Trans. Knowl. Data Eng. **28**(7), 1919–1933 (2016)
37. Wang, Z.Y., Cheng, J.P., Wang, H.X.: Short text understanding: a survey. J Comput. Res. Dev. **53**(2), 262–269 (2016)
38. Wang, Y.S., Huang, H.Y., Feng, C.: Query expansion based on a feedback concept model for microblog retrieval. In: Proceeding of the 26th International Conference on World Wide Web, pp.559–568. ACM (2017)
39. Wiemer-Hastings, P., Wiemer-Hastings, K. and Graesser, A.: Latent Semantic Analysis. In: Proceedings of the 16th international joint conference on Artificial intelligence, pp.1–14. IJCAI (2004)
40. Wu, W.T., Li, H.S., Wang, H.X., Zhu, K.Q.: Probase: a probabilistic taxonomy for text understanding. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, pp.481–492. ACM (2012)
41. Xia, F., Yu, C.C., Xu, L.H., Qian, W.N., Zhou, A.Y.: Top- k Temporal Keyword Search over Social Media Data. World Wide Web. **20**(5), 1–21 (2017)
42. Xu, Z.G., Liu, L., Cai, H.B., Yan, S.Y.: Research on Chinese Microblog's semantic expansion model based on specific context. In: 11th International Conference on International Conference on Fuzzy Systems and Knowledge Discovery, pp.610-615. IEEE (2014)
43. Ye, Z., Huang, J.X., Lin, H.F.: Finding a good query-related topic for boosting pseudo-relevance feedback. J. Assoc. Inf. Sci. Technol. **62**(4), 748–760 (2014)
44. Yu, Z., Wang, H.X., Lin, X.M., Wang, M.: Understanding Short Texts through Semantic Enrichment and Hashing. IEEE Trans. Knowl. Data Eng. **28**(2), 566–579 (2016)
45. Zhao, F., Zhu, Y.J., Jin, H., Yang, L.T.: A personalized hashtag recommendation approach using LDA-based topic model in microblog environment[J]. Futur. Gener. Comput. Syst. **65**, 196–206 (2016)
46. Zhou, D., Wu, X., Zhao, W.Y., Lawless, S., Liu, J.X.: Query expansion with enriched user profiles for personalized search utilizing folksonomy data. IEEE Trans. Knowl. Data Eng. **29**(7), 1536–1548 (2017)
47. Zuo, Y., Zhao, J.C., Xu, K.: Word network topic model: a simple but general solution for short and imbalanced texts. Knowl. Inf. Syst. **48**(2), 379–398 (2016)