information

MDPI

Article

Chinese Microblog Topic Detection through POS-Based Semantic Expansion

Lianhong Ding [1], Bin Sun [1] and Peng Shi [2,*]

[1] School of Information, Beijing Wuzi University, Beijing 101149, China; lhdingbwu@sina.com (L.D.); sunbinbwu@163.com (B.S.)
[2] National Center for Materials Service Safety, University of Science and Technology Beijing, Beijing 100083, China
* Correspondence: shipengustb@sina.com

Received: 25 June 2018; Accepted: 8 August 2018; Published: 10 August 2018

Abstract: A microblog is a new type of social media for information publishing, acquiring, and spreading. Finding the significant topics of a microblog is necessary for popularity tracing and public opinion following. This paper puts forward a method to detect topics from Chinese microblogs. Since traditional methods showed low performance on a short text from a microblog, we put forward a topic detection method based on the semantic description of the microblog post. The semantic expansion of the post supplies more information and clues for topic detection. First, semantic features are extracted from a microblog post. Second, the semantic features are expanded according to a thesaurus. Here TongYiCi CiLin is used as the lexical resource to find words with the same meaning. To overcome the polysemy problem, several semantic expansion strategies based on part-of-speech are introduced and compared. Third, an approach to detect topics based on semantic descriptions and an improved incremental clustering algorithm is introduced. A dataset from Sina Weibo is employed to evaluate our method. Experimental results show that our method can bring about better results both for post clustering and topic detection in Chinese microblogs. We also found that the semantic expansion of nouns is far more efficient than for other parts of speech. The potential mechanism of the phenomenon is also analyzed and discussed.

Keywords: Chinese microblogs; semantic expansion; short text; topic detection

1. Introduction

Nowadays, microblogging services such as Twitter, Google+, Sina Weibo, and Tencent have become important for catching up on news and exchanging information. Since Chinese is one of the most popular global languages, Chinese microblogs play an important role in social activities. Sina Weibo (weibo.com) was launched in 2009 and has become the largest Chinese microblog system in the world. The 2017 financial report of Sina (NASDAQ: SINA) showed that the monthly active users of Weibo had increased to 392 million by the end of 2017. Another large Chinese microblog network is Tencent (0700.HK), whose microblog function is bound to two popular social chat tools, QQ and WeChat, with a total of 1.59 million monthly active users by the end of 2017.

Since a large number of microblog users publish messages at any one time, microblogs have become an important way to spread topics and opinions. Topic detection and tracking (TDT) is a classic problem for natural language processing. The task is to identify new topics and follow existing topics from the information source [1]. Unfortunately, traditional TDT tools work badly on microblog analysis because of the shorter length and user-generated content (UGC) characteristics. For topic detection and clustering of posts in Chinese microblogs, there are several challenges to face [2].

(1) Short Texts. Since microblogs were initially invented for cell phones, general microblog systems place limitations on the length of a post. The length of post published on Twitter was required to be no more than 140 characters. Although Twitter and Weibo have since increased the length limitation to 280 characters and 2000 Chinese characters, respectively, most posts on microblogs are fairly short because of the users' habits.

(2) Scattered Topics. The contents of microblog posts involve personal life, events, and even reposts from others. Thus, the topics of a microblog are quite scattered in different domains, even when authored by just one user.

(3) Sparse Data. The frequently-used formal Chinese words number more than 50,000. Totally new words often emerge from the Internet. The large number of words plus the length limitation of microblog posts create data sparseness. The sparsity problem prevents general data mining methods from achieving the desired accuracy.

(4) Informal Language. As a kind of UGC, a microblog post is colloquial. Abbreviations, slang expressions, and even typos and grammar mistakes are often found in microblog texts. It is a challenge to detect the correct topic by normal natural language processing methods.

(5) Rapid Generation. Due to a great number of users and convenient mobile publishing, the content of Chinese microblogs is updated very quickly and a lot of posts are published every day. One strategy is to treat the messages of a microblog as a kind of continuous data stream.

Obviously, the characteristics of Chinese microblogs cause difficulties for topic detection and clustering by traditional TDT methods. To solve these problems, two major issues should be solved. One is the proper representation method of posts. The other is the selection of clustering algorithms.

Incorporating semantic knowledge has proven to be useful for enhancing text representation [3]. Since short texts such as microblog posts cannot provide enough words for text classification and effective similarity computing, some kinds of expansions are executed. Adding semantics to a short text is a mainstream approach. There is little research about modeling the semantics of individual microblog posts [4]. Some efforts have focused on the use of external knowledge resources. A method introduced in [5] used a search engine, such as Google, to get more contextual information for measuring the similarity of short text snippets. However, the method not only consumes a lot of time but also depends largely on the quality of the search engine. A three-layer architecture was put forward to enrich the representation of features through Wikipedia and Wordnet [6]. Phan derived topics from an external corpus to enhance the characterization of short texts [7]. Quan used dataset topics to build associations between different words directly [8]. In order to enhance the clustering of microblogs, Hu et al. presented a method to enrich text representation with the power of semantic knowledge bases, such as Wikipedia and WordNet [9]. A method to enrich the representation of short texts by Wikipedia was also put forward by Banerjee et al. for Web text applications [10].

In order to measure the similarity between sentences, Amir et al. extracted semantic kernels such as subject, verb, and object from sentences first [11]. Then the similarity between them was calculated using WordNet as the main linguistic resource and DBpedia as the secondary resource. Finally, machine learning was employed to find the best way to combine semantic similarities between two kernels. Meij et al. put forward a method to add semantics to microblog posts. They converted the task of linking tweets with concepts in Wikipedia into a ranking problem of concepts that are related to the post [4]. Wikipedia was used to solve entity disambiguation for noisy short texts by Shirakawa as well [12].

HowNet is an online knowledge base that unveils the inter-conceptual relations and inter-attribute relations of concepts in Chinese lexicons. It has been widely used in Chinese natural language processing, such as word similarity measurement, information extraction, text categorization, question answering, word sense disambiguation, etc. A similarity measure based on HowNet is put forward by Zhang et al. to find product weakness in Chinese reviews [13]. HowNet is utilized to find infrequent features from the dataset. A sentiment dictionary constructed from Hownet is used by Cao et al. for sentiment analysis in microblogs [14]. TongYiCi CiLin (TYCCL for short) is a Chinese semantic lexicon

like WordNet [15]. It has been employed to resolve word sense disambiguation for Chinese and event similarity computation in texts [16,17].

To overcome the sparsity problem with short texts, a semantic representation of Chinese microblog posts based on TYCCL is introduced. A semantic expansion strategy based on part-of-speech (POS for short) is put forward to solve the problem caused by the phenomenon of polysemy in TYCCL. Since this is a continuous data stream, an incremental clustering algorithm, not a batch learning algorithm, is chosen for post clustering to look for potential topics in microblog [18]. With the potential topic and semantic representation of microblog post in hand, a topic detection method for Chinese microblogs based on semi-supervised learning is proposed. Experiments are conducted on the dataset from Sina Weibo to evaluate our method. Experimental results show that the performance of our method increases both for clustering of posts and for topic detection in Chinese microblogs.

The rest of this paper is organized as follows. Section 2 discusses the materials and methods involved: the semantic representation method of Chinese microblog post, the post clustering method based on a single-pass algorithm, and topic detection based on semi-supervised learning. Section 3 details the experiments and results involved: data collection and preprocessing, evaluation criteria, experiments for microblog clustering algorithm, and experiments for topic detection. Section 4 gives some discussion. Finally, Section 5 summarizes the work of this paper.

2. Materials and Methods

2.1. Semantic Representation of Chinese Microblog Posts

2.1.1. Initial Representation

Different from English, the general first step of natural language processing of Chinese is word segmentation. This is the process of separating a sentence with continuous Chinese characters into several meaningful words. In our method, NLPIR is adopted for Chinese word segmentation. It is a mainstream Chinese word segmentation system and can divide Chinese text into words and POS tagging [19]. Each post can be segmented into Chinese words and each word is tagged with POS by NLPIR. Then, stop words are removed. Because of the sparsity problem for short texts, all the remaining words are reserved as feature words.

The initial representation of a microblog post p is a set of (t_i, w_i). Here, t_i is the feature term after stop words are removed. w_i is the weight of t_i and is calculated by a standard TFIDF function, shown in Equation (1):

$$w_i = TF(t_i, p) \cdot IDF(t_i). \tag{1}$$

Here, the term frequency $TF(t_i, p)$ is the number of times term t_i appears in post p. IDF is the inverse document frequency. $IDF(t_i)$ indicates the relative number of posts in which t_i occurs, defined by Equation (2):

$$IDF(t_i) = \log \frac{|D|}{df(t_i)}. \tag{2}$$

Here, $|D|$ is the total number of posts and $df(t_i)$ is the number of posts in which term t_i occurs. For the incremental clustering algorithm, IDF should be computed in advance according to the existing dataset.

2.1.2. Key Feature Extraction

TFIDF simply filters out words whose weights are relatively low or picks out words whose weights are relatively high. Since there are a lot of features in a long text, the feature extraction has less influence on the representation of a long text. However, for short texts like microblog posts, each feature is important to its representation. In order to enhance the representation for microblog posts, semantic analysis is necessary for feature extraction.

In fact, each post can be regarded as a short narrative, although some factors are missed. From the point of view of linguistics, a narrative has six elements: time, place, character, cause, process, and outcome. Among them, cause, process, and outcome can outline an event. For simplicity and feasibility, cause, process, and outcome are summarized as one event. Microblog systems also provide special symbols for users to indicate the main content directly, called a theme. Theme can help users to get the meaning of a post quickly. For the reasons above, we extract five features from each microblog post: time, place, people, event, and theme.

- Time (with Date)

After Chinese word segmentation, each separated word contains a POS tag. The words with "per ton" or "/tg" tagging are the words whose POS is time. Another time of a microblog post is the time when it was created. This can be obtained directly from the corresponding field in the microblog. The regular expression can be written according to the expression form of date and time. Date can be expressed as yyyyMMdd, yyyy-MM-dd, yyyy/MM/dd or yyyy.MM.dd. Time usually has the form hh:mm:ss.

- Place

The words describing place are tagged with "/ns" or "/nsf" by NLPIR. So we can extract this kind of feature from a text in a microblog directly.

- People

Each microblog user has a nickname. The nickname in a microblog always appears in the form of "@nickname " or "@nickname:". So we extract the words between "@" and the blank space or colon following "@" as the features of the nickname. Features of a nickname may be the words between "@" and ":" following "@" as well. Other names of people may appear in the microblog post. NLPIR specifies tagging of "/nr", "/nr1", "/nr2", "/nrj" or "/nrf" to the words about the name.

- Event

An event feature is a set of words that can describe an event. Different users publish microblogs for different purposes with various styles. An event described in Chinese usually consists of a subject, predicate, and object. A subject may be a noun, adjective nominalization, verb noun, or pronoun. Generally, a predicate is a verb and an object is placed after a transitive verb or preposition. The object may be a noun, pronoun, noun adjective, verb noun, etc. Here, nouns, verbs, and adjectives are extracted from microblog text as event features.

- Theme

When a user wants to point out a clear theme for a post, he/she can mark it with a special symbol. Twitter provides hashtags for the theme. Hashtags are good indicators to detect events and trending topics in microblogs [20]. They have been proven to be useful for microblog retrieval [21] and sentiment analysis [22]. However, Chinese microblogs also employ a pair of square brackets or double angle brackets for the label of theme. Extract theme features from those special symbols can save time and improve the accuracy of feature extraction.

2.1.3. Semantic Representation Based on TYCCL

Our semantic expansion method is based on TYCCL. TYCCL is developed by the Harbin Institute of Technology Center for Information Retrieval [23]. Like HowNet, TYCCL also describes the similarity between Chinese words. In addition, TYCCL focuses more on synonyms. It provides synonym sets for Chinese words. TYCCL organizes words in a five-layer hierarchy structure, shown in Table 1.

Chinese vocabulary is divided into large, medium, and small categories. There are 12 large categories, 97 medium categories, and 1428 small categories. Each small category is divided into a number of word groups according to the proximity of meaning and relativity. Each word group is further divided into a number of word units. The words in one word unit are either synonyms or strongly related [24].

Table 1. Description for coding of TYCCL.

Code Position	Mark	Meaning	Level
1	C	large categories	F1
2	b	medium categories	F2
3	0	small categories	F3
4	2		
5	A	word group	F4
6	0	word unit	F5
7	1		
8	=/#/@	"=" means synonym "#" means related words "@" means independent words	

Large categories and medium categories are marked with capital letters and lower case letters, respectively. Small categories are identified by two decimal numbers. Word group and word unit are specified by an uppercase letter and two decimal numbers. The 8th byte is used to explain the relationship among the words in one word unit. Three symbols can be specified to the 8th byte. "=" indicates that the words in the word unit are synonymous or equal. "#" means the words in the word unit are strongly related. "@" means there is no synonym or relevant word for the word in the word unit. For example, the Chinese word "东南西北" is represented in TYCCL as "Cb02A01 = 东南西北 四方". The symbol "=" indicates "四方" is synonymous with "东南西北". Here, "东南西北" means "east south west north" and "四方" means "everywhere".

Semantic representation is a set of term pair (s_i, w_i). Here, s_i represents a feature word after semantic expansion and w_i means the weight of s_i. In order to distinguish it from the initial representation, s_i will be called a feature item in the rest of the paper. The basic method to add semantics to the presentation of a post is described as follows.

To expand feature words, for each feature word x_i, TYCCL is traversed. If there is a "word unit" that contains x_i and the value of the 8th byte in the coding of TYCCL is "=", feature x_i is replaced by a feature item and described as the coding of the "word unit" in TYCCL. If there is no "word unit" that contains x_i or the 8th byte in the coding of the "word unit" is not "=", x_i will be specified by a unique code. The weight of each feature item is calculated by TFIDF function as well.

Since many Chinese words have different meanings in different contexts, a feature word may be included in more than one "word unit" in TYCCL. For these words, there are three strategies considering the different descriptive ability of different parts of speech. The simplest method is to replace the feature term with all "word units" matched with it. The second method expands feature words except for verbs. The third method only replaces nouns by "word units." The three methods are called full semantic expansion, semantic expansion without verbs, and semantic expansion with nouns, respectively. Including the initial representation without semantic expansion, there are four kinds of representation methods in a microblog post.

For example, a post whose content is "咦！不对！我漏看了最后一句话。。。。好吧，应该没我们什么事了。。。。祝公务员放假愉快。。。。" is represented as a set of eight feature words through word segmentation and removing stop words. The eight feature words are "漏", "句", "话", "事", "祝", "公务员", "放假", and "愉快", respectively. Through the TFIDF computing, the initial representation without semantic expansion looks like {("漏", 0.898395716969), ("句", 0.494709966471),

(“话”, 0.457233543932), (“事”, 0.474012276808), (“祝”, 0.707638679031), (“公务员”, 0.317906965181), (“放假”, 0.712722558999), and (“愉快”, 0.815990013359)}.

The feature items through full semantic expansion are “Bp31C01=”, “He15B01=”, “Id20D01=”, “Dk06B01=”, “Dk11A01=”, “Da01A01=”, “Da01C01=”, “Di19A01=”, “Di22A01=”, “Hi37B01=”, “Hj19A01=”, “Hi10A01=”, “Ae01B02=”, “Ga01A01=”, and “0x000ca=”. Among them, the first three items are the coding of the three “word unit” in TYCCL that contains the feature word “漏”. The fourth item is the “word unit” matched with “句”. The fifth item matches with “话”. The sixth to 11th are six “word units” that contain “事”. The 12th item matches with “祝”. The 13th matches with “公务员”. The 14th matches with “愉快”. There are no “word unit” that contains “放假” and is coded with an ending symbol of “=”. So the feature word “愉快” is specified by a unique coding “0x000ca=” in the semantic representation of the post. The full semantic representation of this post is {(“Bp31C01=”, 0.898395716969), (“He15B01=”, 0.573833502569), (“Id20D01=”, 0.734292315814), (“Dk06B01=”, 0.477699441466), (“Dk11A01=”, 0.386117478352), (“Da01A01=”, 0.420405701485), (“Da01C01=”, 0.467743933584), (“Di19A01=”, 0.34191762479), (“Di22A01=”, 0.449820891011), (“Hi37B01=”, 0.472760875922), (“Hj19A01=”, 0.401834372944), (“Hi10A01=”, 0.620092750465), (“Ae01B02=”, 0.317906965181), (“Ga01A01=”, 0.335411283489), and (“0x000ca=”, 0.712722558999)}.

For the example above, the number of feature words is eight and the number of feature items through semantic expansion is 15. In some cases, several feature words with the same meaning may be replaced by one feature item. For example, both “岳父” and “老丈人” mean father-in-law. They will be replaced by one item coding as “Ah07B02=”. Therefore, the number of feature items in a microblog post may be less than the number of feature words. We still call this process a semantic expansion, because the representation of the feature item will add descriptive ability for the words with the same meaning. It also increases the similarity between the post and the topic.

2.2. Microblog Clustering and Topic Detection

2.2.1. Microblog Post Clustering Based on the Single-Pass Algorithm

Traditional clustering algorithm like k-means can be regarded as a kind of batch learning algorithm. There are some drawbacks to the k-means clustering algorithm. First, in order to classify the documents into a specified number of clusters, the value of k has to be determined before clustering. The clustering results greatly depend on the initial value of k. Unfortunately, it is very difficult to predict the right number of topics in the microblog. Secondly, the number of posts in a microblog platform grows over time. As a result, the dataset for clustering changes dynamically and the number of topics may change over time as well. K-means cannot handle newly added microblog posts. Therefore, an incremental clustering algorithm is more suitable for topic detection in microblogging data streams. A single-pass algorithm is a typical incremental clustering algorithm. Referring to the fundamental idea of a single-pass algorithm, single-pass clustering has been used for new online event detection and topic detection [25,26]. The similarity between the current document and each cluster is calculated. If the similarity exceeds a certain threshold, the current is merged into the cluster that is the most similar to it. For microblog post clustering, each cluster indicates a topic detected.

The simplest single-pass clustering algorithm can be regarded as a kind of 1NN clustering algorithm. It will be used as the baseline for microblog post clustering and is depicted as the following Algorithm 1.

There are two disadvantages to the single-pass algorithm based on 1NN. One is that the result of clustering depends on the order in which posts are processed. The other is the large amount of time required. Microblog posts are always processed in the order in which they are created. This order is fixed, so the processing sequence is not a problem for microblog topic clustering. The time complexity of the single-pass based on 1NN is $O(n^2)$. Here, n is the number of microblog posts.

In order to reduce the time taken, it can be improved as in Algorithm 2 and is named single-pass clustering based on the dynamic model in the rest of this paper.

Algorithm 1: Single-pass clustering based on 1NN

Step 1. Load the microblog data /* Process the microblog posts serially */
Step 2. Create the first cluster with the first post /* the first post is regarded as a cluster */
Step 3. For each subsequent post
Calculate the cosine similarity between the current post and each clustered post
If the similarity exceeds a specified threshold /* compare the current post with each post that has been clustered */
Add the current post into the cluster of the clustered post/* this cluster contains the post to which the current post is the most similar */
Else
Create a new cluster with the current post /* the current post is regarded as a new cluster */
End for
Step 4. Output all the clusters

Algorithm 2: Single-pass clustering based on dynamic model

Step 1. Load the microblog data /* Process the microblog posts serially */
Step 2. Create the first cluster with the first post and take the representation of the first post as the initial model of the first cluster /* the first post is regarded as a cluster and the representation of the first post is the initial representation of the first cluster.*/
Step 3. For each subsequent post
Calculate the cosine similarity between the current post and each cluster
If the similarity exceeds a specified threshold /* compare the current post with each cluster's model */
Add the current post into the cluster of the clustered post/* the model of this cluster is the most similar to the current post */
Update the model of this cluster according to the current post /* the representation of the cluster is modified by the current post */
Else
Create a new cluster with the current post /* the current post is regarded as a new cluster and its representation works as the initial model of this cluster.*/
End for
Step 4. Output all the clusters

The time complexity of the single-pass based on the dynamic model is O(kn). Here, k is the number of clusters generated by the algorithm and n is the number of microblog posts.

2.2.2. IDF Calculation and Topic Representation

(1) IDF in single-pass clustering

After each word of the whole microblog posts is replaced with the corresponding feature items, TF and IDF are computed. With incremental clustering algorithms, the number of documents continues to grow. The calculation method for corpus-level statistics like IDF must be adjusted. There are two possible approaches to calculate IDF. One is to compute IDF in advance using a corpus in a similar application domain. The other is to recalculate IDF when a new document is processed. The second method can work well only after a sufficient number of documents have been processed [27], so we choose the first way to calculate IDF. The "past" posts collected from Sina Weibo are used as a document set to calculate IDF.

(2) Initial topic representation

According to the process of Algorithm 2, single-pass clustering based on prototype, the initial topic representation is the representation of the first microblog post that belongs to the topic. In other words, the topic directly adopts the representation of the corresponding post as its initial representation. As described in Section 2.1.3, if a microblog post is described without semantic expansion, it is

represented as a set of feature words and their weights. If semantics are added to microblog posts through TYCCL, a post is represented as a set of feature items and their weights. So a topic (cluster) can be described as a set of feature words and their weights or a set of feature items and their weights as well.

(3) Topic evolution

During the process of microblog post clustering, current posts will be clustered into an existing topic (cluster) or a newly added topic (cluster). When a new post is assigned to a topic (cluster), the topic representation should be updated to reflect the influence from the current post, called topic evolution.

There are different methods to update the representation of a topic. In order to reduce the time taken, a simple method is employed in this paper. It adds the weight of a feature word or feature item in the representation of the post to the weight of the same feature word or feature item in the representation of the topic.

2.2.3. Topic Tracking and Detection Based a Joint of Classification and Clustering

After the clustering of microblogs, there are some clustered topics with stable representation. Regarding the stable topics as classes, we classify new microblogs into different topics. First, a threshold value ε is specified for classification. When a new microblog post appears, its similarity to existing topics is assessed one at a time. The post will be classified into the most similar topic if the similarity is bigger than ε. The topic representation is not updated in the process of semi-supervised learning. For a post that cannot be classified into any existing topics, meaning that no similarity value is bigger than ε, it is reserved for another post clustering procedure for new topic detection.

This process is a joint one of classification and clustering. The clustering results of microblog posts supply a relatively stable template for the process of post classification. On the one hand, the classification process means new posts are classified into existing topics with higher probability. On the other hand, new topics can be detected continuously in the following clustering.

3. Experiments and Results

3.1. Data Collection and Preprocessing

We conduct experiments with a dataset from Sina Weibo. In general, Microblog Open Platform's API (application programming interface) and Web crawler are two common ways to collect data from microblogs. Sina Weibo also provides open APIs for application development. For some business reasons, Sina Weibo modified some APIs in June 2013. More restrictions were imposed on the calling of retrieval functions. On the one hand, the restrictions of APIs such as requests per hour seriously hinder the speed of data retrieval. On the other hand, traditional Web crawlers do not work well because only a logged-in user can see the complete information. So we combine the simulated login technology into Web crawler. Our crawler collects data from Sina Weibo as follows:

(1) Multiple accounts are registered in Sina Weibo and a number of users are specified as seeds manually.

(2) In order to realize virtual login, related packets are sent to the server and a server session is established.

(3) All cookie contents returned by the server are encapsulated in an HTTP package. These contents may be useful for the next simulated login.

(4) After successful log-in, a microblog post is crawled like a common Web page. The related information is extracted and stored in a local database, such as user information, comment number, forwarding number, and so on.

(5) Change user periodically to cope with the anti-crawl mechanism of Sina Weibo. Each user is allowed to visit no more than 30 pages per hour and each page can list no more than 20 microblogs.

In other words, one user can get no more than 600 microblogs each hour. With users changing periodically, our crawler can get information quicker from Sina Weibo.

Table 2 illustrates the crawling results from Sina Weibo. The dataset includes more than 79,000 microblog posts involving 13 topics. They are annotated manually. The five topics receiving the most attention are selected for our experiments. They are "公务员" (civil servant), "同桌的你" (my old classmate), "转基因" (transgenosis), "雾霾" (smog), and "魅族" (Meizu). There are 7300 posts about civil servants, 10,264 about my old classmate, 5388 about transgenosis, 5647 about smog, and 3122 about Meizu. Two thousand posts for each topic are extracted from the dataset randomly, 1000 for the topic clustering experiment and 1000 for the topic detecting experiment.

Table 2. An example of a dataset extracted from Sina Weibo.

Created at	Repost	Comment	Like	Text
2017/5/16 22:14	2	2	0	如今的雾霾天让我更加喜欢看老照片，因为那里有蓝天.....
2017/5/10 23:12	0	5	6	最卡的手机排行，魅族排第二都没人敢排第一，现在就想翻出我的诺基亚砸烂手上的破烂玩意儿！（即将面临王者禁赛的我气得瑟瑟发抖
2017/5/8 21:03	0	0	0	如果记忆可以重写。我希望当初坐在我身边的还是你。——《同桌的你》
2017/5/8 21:31	1	12	13	其实吧，在小城市，考一个公务员编制是拉开幸福生活之门的关键，是解决生活绝大部分问题之源。考上，就有源源不断的优质相亲对象，抛开自身层次不说，对象的层次，就足以保证之后生活质量。我周围例子全部（没有夸张，对，就是全部）证明了这一点。
2017/5/9 21:07	2	1	0	哈哈读书时禁止恋爱 毕业立马要结婚 你以为对象是馅饼啊想掉就掉
2017/4/12 23:40	0	1	2	开学/放假背着一大坨书来回飞。。。搞得好像很爱学习的样子 呵呵

3.2. Evaluation Criteria

The methods to evaluate the effectiveness of topic detection and clustering are the same. They are precision (p), recall (r), F1 measure (F1), miss rate (m), false rate (fa) and cost function ($Cost$), shown in Equations (3) to (8), respectively. F1 measure is a combination of precision and recall. The $Cost$ function combines both false rate and miss rate. The higher F1 is, the better the performance is. The lower $Cost$ is, the better the performance is:

$$p = TP/(TP + FP) \tag{3}$$

$$r = TP/(TP + FN). \tag{4}$$

For a topic t, TP (short for true positive) means the number of posts belonging to t and being successfully classified into the cluster corresponding to t. FP (short for false positive) is the number of posts that are classified into the cluster corresponding to t but do not belong to t. FN (short for false negative) is the number of posts that belong to t but are not part of the cluster that corresponds to topic t. TN (short for true negative) is another correct number that represents the number of posts not belonging to t and not having been assigned to the cluster corresponding to topic t.

$$F1 = (2 * p * r)/(p + r) \tag{5}$$

$$m = FN/(TP + FN) \tag{6}$$

$$f_a = FN/(TN + FN) \tag{7}$$

$$\mathrm{Cost} = C_{f_\alpha} * f_\alpha * (1 - P_{\mathrm{target}}) + C_m * m * P_{\mathrm{target}} \tag{8}$$

Here, C_m is the cost function of miss rate, $C_{f\alpha}$ is the cost function of false rate, and P_{target} is the priori target probability that the post belongs to the topic. In the second phase of NIST's Topic Detection and Tracking research project (TDT2), the cost function was defined with P_{target} = 0.02 and $C_m = C_{f\alpha}$ = 1.0 [25].

3.3. Experiments for Microblog Clustering Algorithm

There are four kinds of representation for a microblog post introduced in Section 2.1.3. In order to evaluate the contribution of semantic expansion to the microblog clustering or topic detection, the single-pass based on prototype is executed when a post is represented in different ways. If a post is represented without semantic expansion, we denote the single-pass based on a prototype as SP-I. When all the feature words are replaced by "word units" according to TYCCL, it is called SP-SF. When verbs in feature words are not expanded, the single-pass based on prototype is called SP-NV. When only nouns in feature words are expanded, it is called SP-ON. We also test the single-pass on 1NN when a post is described without semantic expansion. It is called SP-1NN in the rest of the paper. One thousand microblog posts for each topic are input to Algorithm 2. The largest five clusters output by Algorithm 2 correspond with the five topics detected by Algorithm 2. When the threshold ε is 0.035, the clustering results are fairly good with high F1 and the lowest cost. So Table 3 illustrates the topic clustering results and the measures when the similarity threshold ε adopted in Algorithm 2 is 0.035.

Table 3. The measures of different clustering algorithms on clustered topics when the similarity threshold is 0.035.

Measure	Cluster Algorithm	Topic					Average
		公务员 (Civil Servant)	同桌的你 (My Old Classmate)	转基因 (Transgenosis)	雾霾 (Smog)	魅族 (Meizu, a Popular Cell Phone)	
p	SP-1NN	0.630	0.561	0.690	0.357	0.448	0.537
	SP-I	1.000	0.837	0.965	0.966	0.946	0.943
	SP-SF	0.872	0.988	0.572	0.817	0.438	0.737
	SP-NV	0.902	0.979	0.720	0.942	0.907	0.890
	SP-ON	0.957	0.996	0.987	0.927	0.964	0.966
r	SP-1NN	0.435	0.674	0.318	0.287	0.400	0.423
	SP-I	0.347	0.554	0.529	0.309	0.406	0.429
	SP-SF	0.368	0.675	0.290	0.156	0.140	0.326
	SP-NV	0.398	0.640	0.339	0.26	0.235	0.374
	SP-ON	0.421	0.719	0.532	0.419	0.424	0.503
F1	SP-1NN	0.514	0.612	0.435	0.318	0.423	0.461
	SP-I	0.515	0.667	0.683	0.468	0.568	0.580
	SP-SF	0.518	0.802	0.385	0.262	0.212	0.436
	SP-NV	0.552	0.774	0.461	0.408	0.373	0.514
	SP-ON	0.585	0.835	0.691	0.577	0.589	0.655
m	SP-1NN	0.565	0.326	0.682	0.713	0.600	0.577
	SP-I	0.653	0.446	0.471	0.691	0.594	0.571
	SP-SF	0.632	0.325	0.710	0.844	0.860	0.674
	SP-NV	0.602	0.360	0.661	0.740	0.765	0.626
	SP-ON	0.579	0.281	0.468	0.581	0.576	0.497
f_a	SP-1NN	0.064	0.132	0.036	0.130	0.123	0.097
	SP-I	0.000	0.027	0.005	0.003	0.006	0.008
	SP-SF	0.014	0.002	0.054	0.009	0.045	0.025
	SP-NV	0.011	0.004	0.033	0.004	0.006	0.011
	SP-ON	0.005	0.001	0.002	0.008	0.004	0.004
Cost	SP-1NN	0.018	0.019	0.017	0.027	0.024	0.021
	SP-I	0.013	0.012	0.010	0.014	0.012	0.012
	SP-SF	0.014	0.007	0.020	0.018	0.022	0.016
	SP-NV	0.013	0.008	0.016	0.015	0.016	0.014
	SP-ON	0.012	0.006	0.010	0.012	0.012	0.010

Higher F1 means better performance and higher cost function value means worse performance. According to the average value of F1 and cost function for SP-1NN and SP-I in Table 3, we see that single-pass based on prototype clustering algorithm is better than single-pass based on 1NN no matter the F1 measure or cost function. As introduced in Section 2.2.1, the time complexity of SP-1NN is $O(n^2)$ and the time complexity of SP_IN is $O(kn)$. Our improvement to SP-1NN cuts down the time complexity. At the same time, the efficiency is improved. The average F1 measure of SP-1NN is 0.461. The average F1 measure of SP_I is 0.580. The average F1 measure is improved by more than 25%.

Figure 1 illustrates the running performance of different clustering algorithms through F1 measure. Figure 2 shows the running cost by a cost function defined in TDT2. It can be seen that SP-ON has the best performance. SP-ON has the highest F1 measure and the lowest cost. This indicates that the TYCCL semantic extension on the nouns of feature words has a good performance. The reason is that the representation ability of feature words is increased by the extension, which makes it easy to cluster similar microblogs into an extended topic.

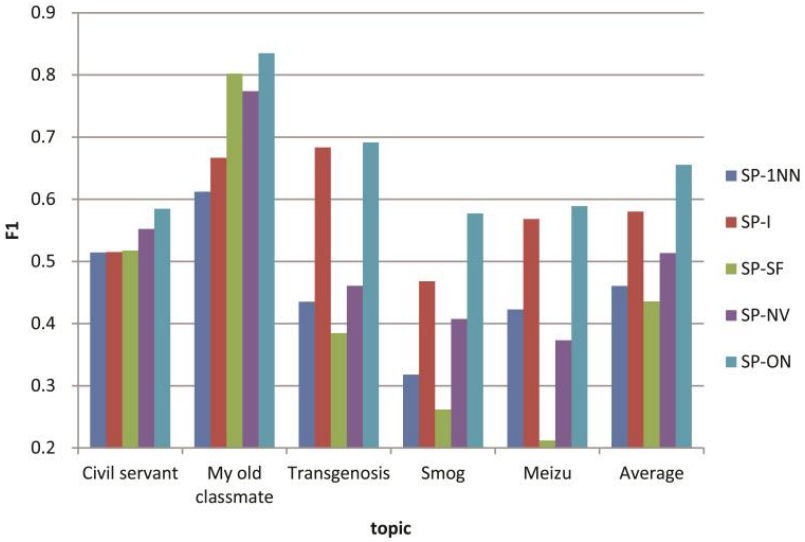

Figure 1. Running performance of post clustering (the threshold is 0.035).

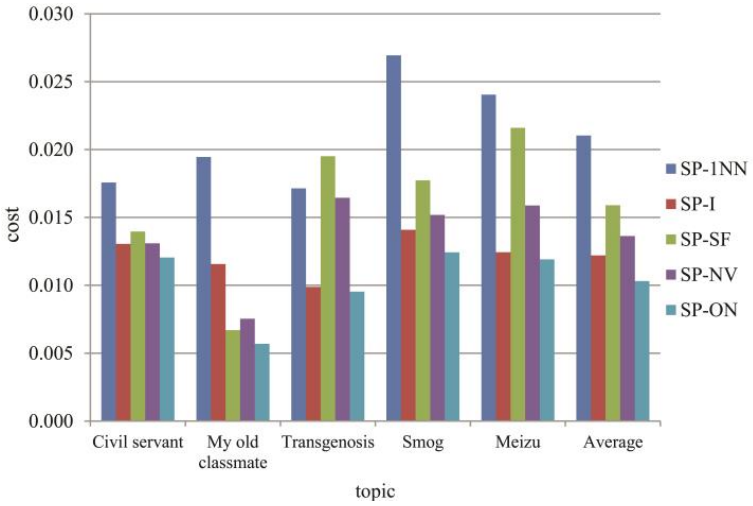

Figure 2. Running cost of post clustering (the threshold is 0.035).

The execution efficiency of SP-I is better than that of SP-SF and SP-NV. This indicates that the TYCCL semantic extension of all the feature words does not improve the clustering precision, but

decreases the efficiency. The extension of all the feature words except for verbs shows the same phenomenon. This could be caused by the frequent phenomenon of polysemy in Chinese. Of the total of 45,365 atomic items of TYCCL, 10,479 of them (23%) have polysemous characterization. Semantic expansion based on TYCCL may solve the data sparseness problem but introduces noise as well. The performance of SP-SF and SP-NV shows that the negative effect of introduced noise exceeds the enhancement of data sparseness when all feature words are expanded or feature words except verbs are expanded. All evaluation criteria for SP-ON are the best. This reveals that the semantic expansion of nouns can improve the representation of microblog posts and enhance the clustering of posts, although there is noise.

Although TYCCL does not tag the POS, the algorithm of Chinese word segmentation gives the tag of POS of every separated word according to the context. In addition, the topic representation ability of nouns is obviously better than that of verbs. Therefore, the nouns in a topic can be extended by TYCCL to improve the topic's representation ability. The verbs and other words in a topic are not extended to preserve the execution efficiency. The performance of SP-ON verifies the strong representation ability of nouns.

To investigate the influence of different similarity thresholds on the clustering methods, several experiments are conducted and the results are shown in Figure 3. It can be seen that the avg_F1 of each method increases with a decrease in the similarity threshold when the similarity threshold is bigger than 0.02. After that, the avg_F1 of SP-SF decreases with the increase in the similarity threshold. A similar phenomenon occurs with SP-I and SP-NV when the similarity threshold is smaller than 0.01. Only avg_F1 of SP-ON continues the trend of increasing with all the similarity threshold values. Its avg_F1 is higher than that of other methods. This shows that the extension of nouns by TYCCL can improve the clustering efficiency of microblog topics.

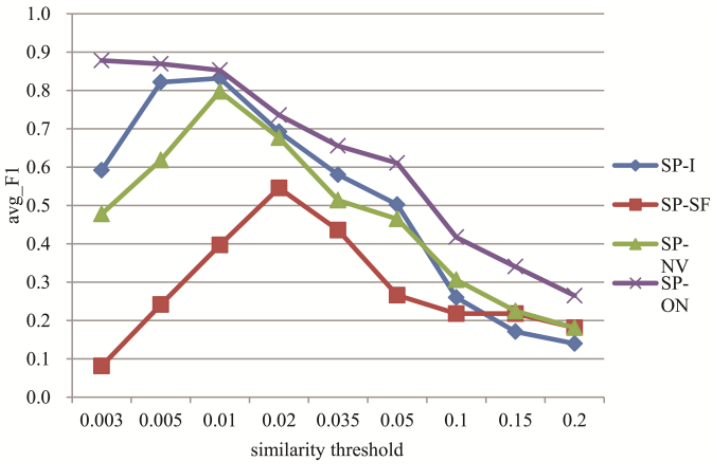

Figure 3. Average F1 measure of microblog clustering for different similarity thresholds.

Figure 4 shows the comparison of avg_cost for different methods with the change in similarity threshold. The avg_cost of each method decreases slowly with the decrease of the similarity threshold at first. The avg_cost reaches the lowest value when the threshold is 0.035. Then avg_cost of each method increases rapidly with the decrease of similarity threshold after a certain value. The reason is that an exception occurs during the clustering. Too low a similarity threshold causes the wrong clustering of microblogs.

The experiments show that SP-ON has the best clustering result and the lowest avg_cost. At the same time, it can tolerate a low similarity threshold, that is, an exception occurs only at the lowest similarity threshold. According to the average of F1, the performance of SP-ON is the best, then SP-I, SP-NV, SP-1NN, and SP-SF.

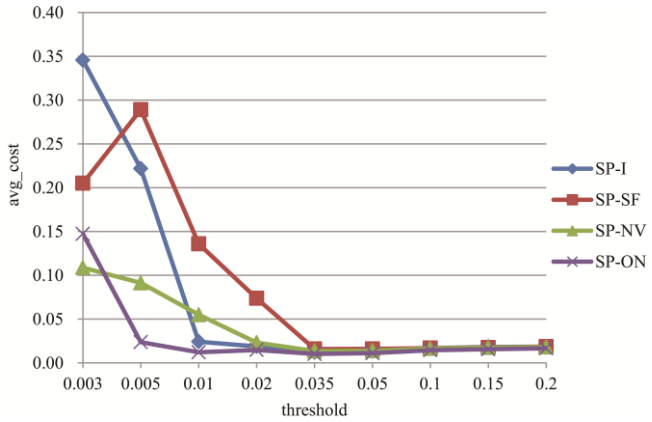

Figure 4. Average cost of microblog clustering for different similarity thresholds.

3.4. Experiments for Topic Tracking

Experiments in this section are conducted according to the process of Section 2.2.3. The process is a joint one of topic tracking and topic detection, but here we focus on the evaluation of topic tracking. Each microblog post will be classified into an existing topic according to the clustering results of Section 3.3 if the similarity is higher than threshold ε. Otherwise, it will be put into another cluster for new topic detection. The cluster representation output when the threshold is 0.035 is applied in this section for topic tracking by classification.

From the comparison above, we see that the clustering result is the best when only the nouns of feature words are expanded. Consequently, only SP-I, SP-SF, and SP-ON are tested for the topic detection on new posts from microblogs.

The topic tracking performance is good when $\varepsilon = 0.15$. Confined by the length of the paper, this section only shows the experimental results when $\varepsilon = 0.15$. Figure 5 shows the performance of topic tracking results using topic representation as a classifying template when the similarity threshold is equal to 0.15. SP-I_r, SP-I_p, and SP-I_F1 are the recall, precision, and F1 measure with the original microblog and topic, respectively. SP-SF_r, SP-SF_p, and SP-SF_F1 are the recall, precision, and F1 measure with the semantic extension of microblog and topic where all words are expanded. SP-ON_r, SP-ON_p, and SP-ON_F1 are the recall, precision, and F1 measure with the semantic extension of microblog and topic where only nouns are expanded. The result shows that the measures with semantic extension of nouns are better than without extension. The performance of topic tracking is the worst when words with any POS are expanded.

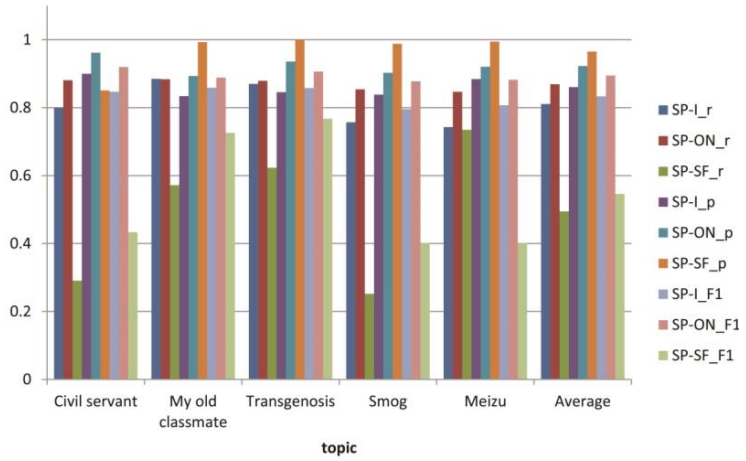

Figure 5. Topic tracking performance when the similarity threshold is equal to 0.15.

Figure 6 shows the cost of topic tracking using topic representation as a classifying template when the similarity threshold is equal to 0.15. SP-I_fa, SP-I_m, and SP-I_Cost are the false, miss, and cost function with the original microblog and topic, respectively. SP-SF_fa, SP-SF_m, and SP-SF_Cost are the false, miss, and cost function with the semantic extension of microblog and topic where all words are expanded. SP-ON_fa, SP-ON_m, and SP-ON_Cost are the false, miss, and cost function with semantic extension of nouns, respectively. SP-ON has lower cost than SP-I. It is verified that the semantic extension of nouns can improve the efficiency of the algorithm operation.

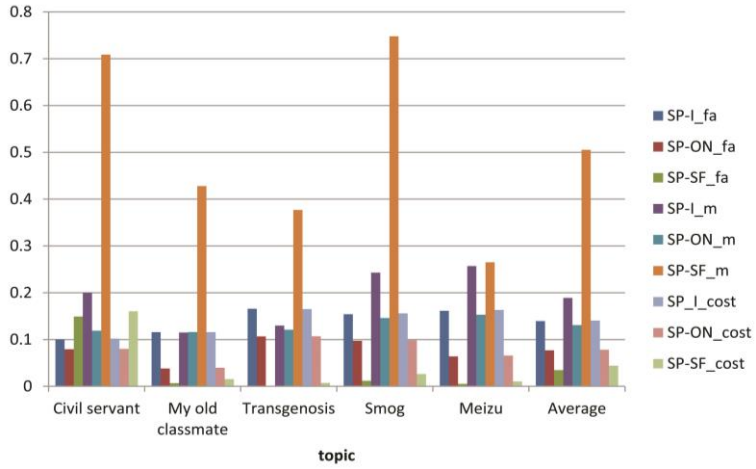

Figure 6. Topic tracking cost when similarity threshold is equal to 0.15.

Table 4 gives the result comparison of topic tracking when the similarity threshold is equal to 0.15. It is obvious that the recall, precision, and F1 measure of semantic extension about nouns is higher than those with no extension. At the same time, the false rate, miss rate, and cost function of semantic extension about nouns is lower than those with no extension. Semantic expansion of nouns by TYCCL can improve the precision and recall of topic detection. At the same time, it can cut down the false rate and miss rate.

Table 4. Topic tracking result when the similarity threshold is 0.15.

Evaluation Criterion	Algorithm	Topic					Average
		公务员 (Civil Servant)	魅族 (Meizu)	转基因 (Transgenosis)	雾霾 (Smog)	同桌的你 (my Old Classmate)	
Recall	SP-I	0.800	0.743	0.87	0.757	0.885	0.811
	SP-ON	0.881	0.847	0.879	0.854	0.884	0.869
Precision	SP-I	0.900	0.884	0.846	0.839	0.834	0.861
	SP-ON	0.921	0.962	0.893	0.903	0.936	0.923
F1-measure	SP-I	0.847	0.807	0.858	0.796	0.859	0.833
	SP-ON	0.920	0.882	0.907	0.878	0.889	0.895
False	SP-I	0.100	0.161	0.166	0.154	0.116	0.139
	SP-ON	0.079	0.064	0.107	0.097	0.038	0.077
Miss	SP-I	0.200	0.257	0.13	0.243	0.115	0.189
	SP-ON	0.119	0.153	0.121	0.146	0.116	0.131
Cost function	SP-I	0.102	0.163	0.165	0.156	0.116	0.14
	SP-ON	0.080	0.066	0.107	0.098	0.040	0.078

Besides replacing feature words with the corresponding feature items, introduced in Section 2.1.3, there is another way to achieve the semantic expansion of a microblog post. It is a joint bag-of-word representation including both feature words and feature items. We name the first way the item method and the second way the joint method, respectively. The average recall, average precision, and average

F1 of topic tracking are illustrated in Figure 7. In the legend, "Item-SF" indicates that words with any POS tag are expanded by the item method. "Joint-SF" means that words with any POS tag are expanded by the joint method. "Item-ON" indicates that only nouns are expanded by the item method. "Joint-ON" means that only nouns are expanded by the joint method. From Figure 7, it can be seen that the average recall and the average F1 of the item method are higher than those of the joint method.

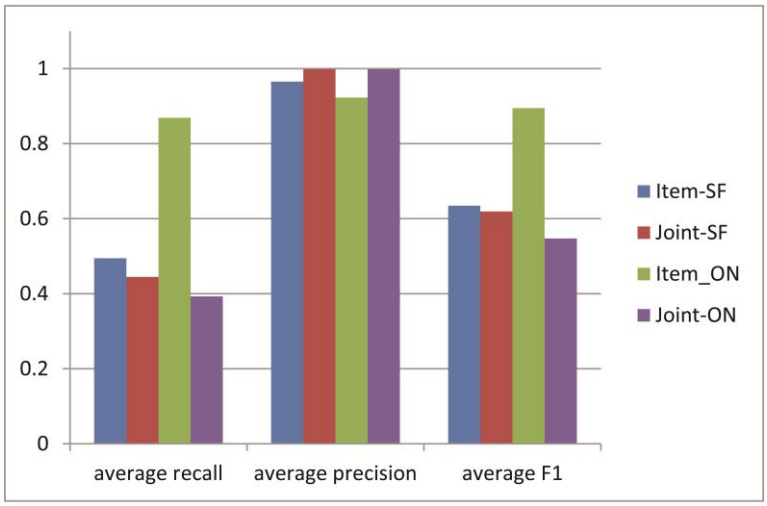

Figure 7. Average recall, precision, and F1 when the similarity threshold is equal to 0.15.

4. Discussion

In order to overcome the data sparsity problem of short texts, a popular Chinese semantic lexicon, TYCCL, is introduced for the process of semantic representation of posts. Several semantic expansion strategies of microblog posts based on TYCCL are compared by experiments. When all the feature words are expanded, SP-I works poorly. The reason may be the existence of polysemous words. The experiment conducted by Liu indicates that the expansion of polysemous words weakens the Chinese entity relation extraction [28]. Many statistical methods have been put forward for the identification of polysemous words. Corpus-based approaches are often used. Corpus-based approaches identify the word sense according to the co-occurrence frequencies extracted from large textual corpora. These approaches have the advantages of flexibility and generality but suffer from a knowledge acquisition bottleneck [29]. In this paper, several semantic expansions for microblog posts are conducted according to the POS of feature words. They are adding semantics for all feature words, adding semantics for feature words whose POS is not a verb, and adding semantics for feature words whose POS is a noun, respectively. SP-I works poorly when semantics are added to all the feature words or semantics are added to the feature words whose POS is not a verb. SP-I works best when only semantics are added to nouns in feature words.

In the paper, every word gets a POS tag through Chinese word segmentation according to the context. The semantic expansion of posts according to POS does not require more computing. The expansion of a noun according to TYCCL can improve the clustering efficiency of posts and the detection of the topic. This, in turn, demonstrates that nouns have more descriptive ability for Chinese microblog posts and topics.

5. Conclusions

In this paper, we propose a Chinese microblog topic detection method based on the improvement of single-pass clustering algorithm and semantic representation of microblog posts. Firstly, Chinese microblog posts without semantic expansion are clustered by single-pass on 1NN and single-pass based on prototype, respectively. The single-pass based on prototype performs better than single-pass

on 1NN. SP-I has higher average precision, recall, and F1 measure and lower average miss rate, false rate, and cost function. At the same time, SP-I has lower time complexity as well. Also, the improved single-pass clustering is more efficient and lower in cost than the original single-pass clustering.

Author Contributions: L.D. and P.S. conceived and designed the methodology and the experiments; B.S. performed the experiments and analyzed the data; L.D. wrote the paper. All the authors have read and approved the final manuscript.

Funding: This research received no external funding.

Acknowledgments: This work was supported by the Beijing Intelligent Logistics System Collaborative Innovation Center, a Breeding Project of BWU (No. GJB20162002), the National Key R&D Program of China (No. 2017YFB0803302), and the Beijing Social Science Foundation (No. 17GLC066).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Castellanos, A.; Cigarrn, J.; Garca-Serrano, A. Formal concept analysis for topic detection. *Inf. Syst.* **2017**, *66*, 24–42. [CrossRef]
2. Wu, F.; Song, Y.; Huang, Y. Microblog sentiment classification with heterogeneous sentiment knowledge. *Inf. Sci.* **2016**, *373*, 149–164. [CrossRef]
3. Hu, J.; Fang, L.; Cao, Y.; Zeng, H.-J.; Li, H.; Yang, Q.; Chen, Z. Enhancing text clustering by leveraging Wikipedia semantics. In Proceedings of the 31st Annual International ACM SIGIR Conference Research and Development in Information Retrieval, Singapore, 20–24 July 2008; pp. 179–186.
4. Meij, E.; Weerkamp, W.; Rijke, M.D. Adding semantics to microblog posts. In Proceedings of the 12th Conference WSDM, Seattle, WA, USA, 8–12 February 2012; pp. 563–572.
5. Sahami, M.; Heilman, T.D. A web-based kernel function for measuring the similarity of short text snippets. In Proceedings of the ACM International Conference World Wide Web, Edinburgh, UK, 22–26 May 2006; pp. 377–386.
6. Hu, X.; Sun, N.; Zhang, C. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In Proceedings of the ACM International Conference Information and knowledge management, Hong Kong, China, 2–6 November 2009; pp. 919–928.
7. Phan, X.H.; Nguyen, L.M.; Horiguchi, S. Learning to classify short and sparse text & web with hidden topics from largescale data collections. In Proceedings of the 17th ACM International Conference World Wide Web, Beijing, China, 21–25 April 2008; pp. 91–100.
8. Quan, X.; Liu, G.; Lu, Z.; Ni, X.; Liu, W. Short text similarity based on probabilistic topics. *Knowl. Inf. Syst.* **2010**, *25*, 473–491. [CrossRef]
9. Hu, X.; Tang, L.; Liu, H. Embracing information explosion without choking: Clustering and labeling in microblogging. *IEEE Trans. Big Data* **2015**, *1*, 35–46. [CrossRef]
10. Banerjee, S.; Ramanathan, K.; Gupta, A. Clustering short texts using Wikipedia. In Proceedings of the 30th Annual International ACM SIGIR Conference Research and Development in Information Retrieval, Amsterdam, The Netherlands, 23–27 July 2007; pp. 787–788.
11. Amir, S.; Tanasescu, A.; Zighed, D.A. Sentence similarity based on semantic kernels for intelligent text retrieval. *J. Intell. Inf. Syst.* **2017**, *48*, 675–689. [CrossRef]
12. Shirakawa, M.; Nakayama, K.; Hara, T. Wikipedia-based semantic similarity measurements for noisy short texts using extended naive bayes. *IEEE Trans. Emerg. Top. Comput.* **2015**, *3*, 205–219. [CrossRef]
13. Zhang, W.H.; Xu, H.; Wan, W. Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis. *Expert Syst. Appl.* **2012**, *39*, 10283–10291. [CrossRef]
14. Cao, D.; Ji, R.; Lin, D. A cross-media public sentiment analysis system for microblog. *Multimed. Syst.* **2016**, *22*, 479–486. [CrossRef]
15. Fellbaum, C. *WordNet: An. Electronic Lexical Database*; MIT Press: Boston, MA, USA, 1998.
16. Lu, Z.; Liu, Y.; Zhao, S.; Chen, X. Study on feature selection and weighting based on synonym merge in text categorization. In Proceedings of the IEEE International Conference Future Networks, Hainan, China, 22–24 January 2010; pp. 105–109.

17. Zhang, X.; Liu, Z.; Liu, W. Event similarity computation in text. In Proceedings of the IEEE International Conference Internet of Things, and Cyber, Physical and Social Computing, Dalian, China, 19–22 October 2011; pp. 419–423.
18. Li, L.; Ye, J.; Deng, F.; Xiong, S.; Zhong, L. A comparison study of clustering algorithms for microblog posts. *Cluster Comput.* **2016**, *19*, 1333–1345. [CrossRef]
19. Zhou, L.; Zhang, D. NLPIR: A theoretical framework for applying natural language processing to information retrieval. *J. Am. Soc. Inf. Sci. Technol.* **2003**, *54*, 115–123. [CrossRef]
20. Kwak, H.; Lee, C.; Park, H.; Moon, S. What is Twitter, a social network or a news media? In Proceedings of the ACM International Conference World Wide Web, Raleigh, NC, USA, 26–30 April 2010; pp. 591–600.
21. Efron, M. Hashtag retrieval in a microblogging environment. In Proceedings of the 33th Annual International ACM SIGIR Conference Research and Development in Information Retrieval, SIGIR '10, Geneva, Switzerland, 19–23 July 2010; pp. 787–788.
22. Davidov, D.; Tsur, O.; Rappoport, A. Enhanced sentiment learning using twitter hashtags and smileys. In Proceedings of the International Computational Linguistics: Posters, COLING'10, Beijing, China, 23–27 August 2010; pp. 241–249.
23. Tongyici Cilin (Extended). Available online: http://ir.hit.edu.cn/demo/ltp/SharingPlan.htm (accessed on 9 August 2018). (In Chinese)
24. Liu, Y.; Zhu, Y.; Xin, G. Chinese Text watermarking method based on TongYiCi CiLin. *Int. Dig. Cont. Techn. Appl.* **2012**, *6*, 465–473.
25. Papka, R. On-line New Event Detection, Clustering, and Tracking. Ph.D. Thesis, Department of Computer Science, University of Massachusetts Amherst, MA, USA, 1999.
26. Huang, B.; Yang, Y.; Mahmood, A. Microblog topic detection based on LDA model and single-pass clustering. In Proceedings of the International Rough Sets and Current Trends in Computing, Chengdu, China, 17–20 August 2012; pp. 166–171.
27. Yang, Y.; Pierce, T.; Carbonell, J. A study of retrospective and on-line event detection. In Proceedings of the 21st Annual International ACM SIGIR Conference Research and Development in Information Retrieval, Melbourne, Australia, 24–28 August 1998; pp. 28–36.
28. Liu, D.; Peng, C.; Qian, G.; Zhou, G. The effect of TongYiCi CiLin in Chinese entity relation extraction. *J. Chin. Inf. Proc.* **2014**, *28*, 91–99. (In Chinese)
29. Leacock, C.; Miller, G.; Chodorow, M. Using corpus statistics and WordNet relations for sense identification. *Comput. Linguist.* **1998**, *24*, 147–165.