Modeling Chinese Microblogs with Five Ws for Topic Hashtags Extraction

Zhibin Zhao, Jiahong Sun, Lan Yao*, Xun Wang, Jiahong Chu, Huan Liu, and Ge Yu

Abstract: Hashtags are important metadata in microblogs and are used to mark topics or index messages. However, statistics show that hashtags are absent from most microblogs. This poses great challenges for the retrieval and analysis of these tagless microblogs. In this paper, we summarize the similarity between microblogs and short-message-style news, and then propose an algorithm, named 5WTAG, for detecting microblog topics based on a model of five Ws (When, Where, Who, What, how). As five-W attributes are the core components in event description, it is guaranteed theoretically that 5WTAG can properly extract semantic topics from microblogs. We introduce the detailed procedure of the algorithm in this paper including spam microblog identification, microblog segmentation, and candidate hashtag construction. In addition, we propose a novel recommendation computing method for ranking candidate hashtags, which combines syntax and semantic analysis and observes the distribution of artificial topic hashtags. Finally, we conduct comprehensive experiments to verify the semantic correctness and completeness of the candidate hashtags, as well as the accuracy of the recommendation method using real data from Sina Weibo.

Key words: hashtag; microblog; topic detection; short-message-style news; five Ws

1 Introduction

A hashtag is a type of label or metadata tag used on social network and microblogging services. It was first used within IRC networks to label groups and topics[1]. Subsequently, hashtags have been used to mark individual messages as relevant to a particular group and to mark individual messages as belonging to a particular topic or “channel”. According to Twitter, a hashtag, prefixed with the symbol “#”, is a word or acronym used to describe a tweet in order for people to follow a conversation easily[2]. In Sina Weibo, the most popular Chinese microblogging system on the Web, a hashtag is defined more explicitly as a keyword that can represent the topics of a microblog (also called Weibo) and be used for microblog retrieval.

Essentially, as defined by Twitter and Sina Weibo, a hashtag is the topic of a microblog. It has great benefits not only to the microblog users but also to the microblogging service providers, and therefore attracts much attention from the academic world.

On one hand, for microblog users, hashtags help accelerate the retrieval of topic-specific microblogs and promote the precision of search results. Given hashtags, users can easily find and follow conversations of interest. Otherwise, they have to turn to techniques involving full-text searches. Full-text searching is distinct from keyword-based searching (such as titles and topic keywords). In a full-text search, the search engine examines all the words in every stored document as it tries to match search criteria (text specified by a user). Obviously, as far as the result precision
and search efficiency are concerned, full-text search techniques are redundant for massive microblogs.

On the other hand, for microblogging service providers, hashtags help to classify microblogs in a more accurate and efficient way. Fine-detailed topic-oriented classification of microblogs is a significant prerequisite for event detection, sentiment analysis, and public opinion mining.

Despite the great importance of hashtags in microblogging systems, its generation totally relies on the freewill participation of users. To encourage and help users to tag their tweets, Twitter released a detailed user guide\textsuperscript{[2]} to explain how to choose and mark a topic hashtag. Sina Weibo tried to provide a more friendly and convenient interface to guide users to propose a hashtag for every microblog. Unfortunately, these efforts have not led to satisfactory results. Meng et al.\textsuperscript{[3]} measured over 0.2 million tweets and found that only around 23\% had at least one \#hashtag. The situation is worse in Sina Weibo. We measured 840,593 Weiboes and found that only 108,714 of them, as low as 12.9\%, had at least one artificial tag. These statistics show that most microblogs have no topic tags, which inevitably becomes a large obstacle to retrieving them.

In this paper, we focus on the research issue of topic detection and propose a novel algorithm named 5WTAG, which aims to automatically generate and recommend a topic hashtag for each Chinese microblog. As microblogs are quite similar to short-message-style news not only in content but also in structure, we suggest modeling microblogs using the five Ws (5W): When, Where, Who, What, and hoW. These are generally accepted as the essential components of news formation. Based on the 5W model, 5WTAG can produce several candidate topic hashtags. Then, the embedded recommendation computation function evaluates each of the candidate topic hashtags regarding semantics, syntax, and empirical data. Finally, all recommended hashtags are clustered for topic merging.

To sum up, there are three contributions in this paper:

- According to detailed analysis of the similarity between microblogs and short-message-style news in content and structure, we propose modeling microblogs with the five Ws. As answering the 5W problems can describe the event completely and correctly, it is more rational to use the 5W model to express the semantics hidden in microblogs.

- We propose a complete solution to extract possible topic tags from Chinese microblogs. This includes four phases: filtering spam microblogs, handling popular Internet words and sentences, segmenting a microblog into clauses, and finally constructing candidate topic hashtags.

- We introduce a quantitative method for computing the recommendation of a candidate topic tag, in which semantic completeness and correctness, importance of content and location, and statistical distribution of artificial hashtags are all considered. In order to use Affinity Propagation to semantically cluster the hashtags, we suggest a novel 5W-based method to measure the topic similarities of pairs of hashtags.

This paper is organized as follows: Section 2 summarizes the related work in topic detection, particularly some up-to-date research progress on topic-oriented microblog analysis problems. In Section 3, we introduce the 5W model and explain why and how it can be used to model microblogs. In Sections 4 and 5, we describe in detail the procedure to construct candidate topic tags for each microblog and the quantitative method of recommendation computation, respectively. We report our performance experiments in Section 6 and conclude our work in Section 7.

2 Related Work

Topic detection is always a hot topic in the academic community. Most of the initial studies focused on text streams or massive document sets and mainly used the technique of document clustering\textsuperscript{[4, 5], Document clustering has two primary technical issues: (1) how the documents are modeled, and (2) how the similarities among the documents are measured. These two issues are closely related and have effect on the clustering results as well as the topic detection results. Zhang and Wang\textsuperscript{[6]} summarized several classical document representation models and similarity calculation methods.

In recent years, popular topic models, such as probabilistic Latent Semantic Analysis (pLSA)\textsuperscript{[7]} and Latent Dirichlet Allocation (LDA)\textsuperscript{[8]}, have received extensive attention. Xu and Wang\textsuperscript{[9]} introduced topic models and the similarity calculated with these models in detail. Essentially, topic models are statistical models and depend on the basic assumption that words relevant to the topics in an article appear more frequently than others. Topic models perform better in conventional documents, but worse in short messages (such as microblogs) due to the severe sparsity of word co-occurrence patterns. Therefore, researchers
have attempted to aggregate short texts into pseudo-documents with some heuristic strategies to enrich context information. Such strategies are generally based on authorship, shared words, word relations, hashtags, and topic inference. For example, the Biterm Topic Model (BTM) learns topics by directly modeling the generation of word co-occurrence patterns in the whole corpus, making the inference effective by using the rich corpus-level information. However, as microblog posts have various interests and concerns, they have little connection with each other, which inevitably poses great challenges to an aggregation method. More recently, Li et al. proposed organizing microblogs as conversation trees by using the repost or reply relationship; this explicitly exploits the topic dependencies contained in conversation structures for microblog aggregation.

Storage the work in this paper belongs to the same research field as that mentioned above, they are quite different in essence: the 5WTAG algorithm aims to detect topics from individual text, such as a piece of a tweet or Sina Weibo. To this end, semantic analysis along with syntactic analysis and empirical statistics, instead of clustering technologies, are used for topic hashtag construction and recommendation calculation.

The huge volume of microblogs carrying sentiments necessitates automatic sentiment analysis techniques, which assist users when summarizing public opinions. Most research work in microblog-oriented sentiment analysis is based on the assumption that microblogs already have hashtags. For example, Meng et al. conducted a comprehensive study on the problem of entity-centric (such as celebrities and brands) topic-oriented opinion summarization on Twitter. To produce opinion summaries and the remarkable insight behind the opinions in accordance with certain entities and topics, they first have to determine the topic of the tweet. Topic extraction in Ref. is carried out using existing human-annotated semantic tags in tweets. This is apparently different from our work as we aim to mine topics from untagged microblogs.

3 Modeling Microblogs with Five Ws

In this section, we start with an explicit definition of the five Ws, which comes from the field of news writing. Subsequently, we summarize the similarity between short-message-style news and microblogs, which proves to be the theoretical foundation for modeling a microblog with the five Ws. At the end of this section, we show how we map each of the words in microblog to an attribute in the five Ws and give a formal description for microblogs using the 5W model.

3.1 Five Ws in journalism

The 5W model has been attributed to Thomas Wilson, who was an English rhetorician and introduced the method in his discussion of the “seven circumstances” of medieval rhetoric. Nowadays, it is often mentioned in journalism and refers to five interrogative words: Who, What, When, Where, and hoW (sometimes six Ws with Why added). They are exactly the reporters’ questions to which the answers are considered basic in information-gathering. To be specific, a report can only be considered complete if it answers the following questions that start with an interrogative word:

- Who did that?
- What was involved?
- When did it take place?
- Where did it take place?
- How did it happen?

Each question above should have a factual answer, i.e., the facts necessary for a report to be considered complete. Importantly, none of these five questions can be answered with a simple “yes” or “no”. So, we define 5W as {“When”, “Where”, “Who”, “What”, “hoW”}.

3.2 Similarities between microblogs and short-message-style news

Since the emergence of microblogging systems, they play the role of new media in human society. This breaks down the greatest barrier between journalists and common people, i.e., press releases can only be delivered by journalists, which is inherent in traditional media such as television, radio, and newspapers. Nowadays, more and more social events are reported by common people with microblogs. Actually, microblogs share many common features with short-message-style news.

3.2.1 Both center on event description

The context-based similarity comparison between blogs and web news has been well studied in recent years. In terms of content, they both center on event description. Generally, short-message-style news concentrates more on public events, while microblogs are more about individual events or social events that happen around the microblog users. As a typical type of “we-media”, microblogs often play the role of a
news report. For example, at local time 8:02 a.m. on April 20, 2013, a strong earthquake struck Ya’an in Sichuan Province. After only 53 seconds, a piece of microblog reporting this event was released via Sina Weibo. In the following hour, over 1000 pieces of microblogs regarding the earthquake appeared on the Sina microblogging system.

Admittedly, some microblogs are just about personal life. For example, someone may release a series of microblogs reporting his or her traveling experience and comments. However, these microblogs can also be regarded as short-message-style news in terms of content. The only difference between these individual-related microblogs and public news is that the former are more attractive to a limited social circle and have far less social influence than the latter.

3.2.2 Both are structured in “inverted pyramid”
The “inverted pyramid”[^25] is a metaphor used by journalists and other writers to illustrate how information should be prioritized and structured in text, especially in a news report. The acknowledged classic example of news with “inverted pyramid” format is the newsflash report about assassination of President John F. Kennedy from Reuters on Nov. 22, 1963. The “inverted pyramid” format is valued primarily for two reasons: First, readers can leave the story at any point and understand it, even if they do not have all the details. Second, for those readers who wish to proceed, it conducts them through the details of the story.

![Fig. 1 The inverted pyramid format.](image)

Investigation shows that if a writer is limited in the number of characters when describing an event (For example, up to 140 Chinese characters in one piece of Sina Weibo), the format of “inverted pyramid” is their most preferred writing style. Therefore, it has adaptability to both short-message-style news and microblogs. People are more inclined to put crucial information at the beginning. In other words, the earlier a piece of information appears in a microblog, the more important it is for event description. The structural characteristics of microblogs provide a significant guideline when distinguishing the level of importance of every part of a microblog.

3.3 Mapping Chinese microblogs to five Ws
According to the analysis above, a microblog can be considered as a piece of short-message-style news. This inspired us to use the 5W model to describe events hidden in a microblog. Let \((w, h)\) represent a piece of a Chinese microblog. \(w\) is its content and \(h\) is its topic hashtag. Next, we introduced how we mapped each of the notional words in \(w\) to one of the five Ws.

First of all, we assumed that Chinese word segmentation tools can distinguish notional words accurately, such as people, institutes, locations, and time. The justification for our assumption is as follows:
- Most mainstream Chinese segmenters already have entity recognition as their inherent functions, such as ICTCLAS2013[^26]. In addition, these word segmentation tools can mark-up a word with its specific Part-Of-Speech (POS) according to not only its definition but also its context. For example, in ICTCLAS2013, a word followed by “/nr” represents a person’s name, while a word followed by “/r” is a pronoun referring to someone.
- As for proper nouns or fixed combination terms (such as Chinese idioms or slang words, abbreviations, and Internet popular words and sentences), we handled them using two methods: First, some Chinese segmenters can recognize these special words. For example, ICTCLAS2013 can mark-up a Chinese idiom with “/i”. Second, all Chinese segmenters can be supplemented with a user-defined dictionary, which helps to promote the segmentation precision and efficiency.

In our implementation of the 5WTAG algorithm, we used ICTCLAS2013 as our Chinese word segmenter and complemented it with a user-defined dictionary. Then, we classified each of the words in \(w\) into one of the five Ws according to the following definitions:
Definition 1 “When” is the set of words that refers to time, festivals, and Chinese solar terms in microblogs.

Definition 2 “Where” is the set of the words that refers to geographical nouns in microblogs.

Definition 3 “Who” is the set of the words that refers to people’s names, a group of people, or institutes. It also includes the personal pronouns in microblogs.

Definition 4 “What” is the set of the words that refers to things or abstract concepts in microblogs. Of particular note, it also includes some proper nouns such as movie or television products, music products, novels, PC games, and commodities and trademarks.

Definition 5 “how” is the set of the words that refers to actions or states of being in microblogs.

Let $X_x$ represent the set of words contained in text $x$ and classified to attribute $X$ in the five Ws. According to Definitions 1 to 5, the 5W model of the microblog $(w, \text{null})$ can be illustrated as $\bigcup_{X \in 5W} \Psi_X(w)$.

We implemented the mapping relation from part-of-speech annotations in ICTCLAS2013 to the attribute $X$ in the five Ws. This is listed in Table 1.

### 4 Extracting Candidate Topic Hashtags

5WTAG is designed to be integrated into microblogging services and run online on the server side. The input for 5WTAG is a piece of microblog without any topic hashtags. The output is a recommended topic hashtag that ranks highest among all the candidate hashtags according to the score of the recommendation computation. The preprocessor is responsible for identifying and removing spam data as well as segmenting the microblogs into multiple clauses. Afterwards, it extracts candidate topic hashtags from each of the clauses. Finally, the recommendation computation module calculates recommendations for each of the candidate hashtags according to several considerations. Figure 2 depicts the overall framework.

#### 4.1 Identifying spam microblogs

As mentioned in Section 1, microblog topic detection is a prerequisite for hot social event identification, public opinion summary, and entity/event-oriented sentiment analysis. Therefore, we should concentrate more on the microblogs with meaningful content. However, a large proportion of “spam” microblogs currently exist. We randomly sampled 1000 pieces of microblogs from our dataset and manually annotated the meaningless microblogs. Table 2 shows our statistical results.

We cannot extract any valuable topics from spam texts. Therefore, we have to filter them out. In

| Type                          | Statistics |
|-------------------------------|------------|
| Event-related microblogs      | 701        |
| Spam microblogs               | 299        |
| Advertisement or redirect URL | 57         |
| Messy code or emotion icon    | 64         |
| Retweets with meaningless comments | 178       |

---

**Table 1** Mapping the part-of-speech annotations to five Ws.

| $X$ | Annotations in ICTCLAS2013 |
|-----|---------------------------|
| When | /t;                      |
| Where | /ns, /nd, /s;             |
| Who | /nr, /nt, /r;            |
| What | /nz, /n (except /ns, /nd, /nr, /nt, /nz); |
| how | /v, /vn;                  |

---

**Table 2** The statistic result of spam microblogs.

Fig. 2 The overall framework of 5WTAG.
our previous research we proposed solutions for the detection and filtering of spam microblogs and these can be integrated into the 5WTAG algorithm. Table 3 lists the methods used in the 5WTAG algorithm to identify the different types of spam microblogs as well as how we tagged them accordingly. The interested readers are directed to Refs. [27, 28] for more details.

We collected words frequently used in advertisements (such as “huigu (visit and buy)”, “dazhe (on sale)”, “tuangou (group buying)”), and further developed a high-frequency word lexicon of advertising. If a microblog contained such words, we concluded that it should be an advertisement.

4.2 Processing Internet catchwords
Internet catchwords are informal words and expressions that are not considered standard from the perspective of Chinese grammar or dialect but are understandable in certain social settings. Since Internet catchwords influence topic recognition, we considered them first.

4.2.1 Popular Internet words
In general, popular Internet words come from two source types: First, they are existing Chinese words that are given new implications. For example, “dajiangyou” previously meant to buy soy sauce, but now it implies that someone is hanging around or not involved in an event. Similarly, “shanzhai” meant a cottage on a mountain, but now it is another way of saying “counterfeiting” with sarcasm. Second, they are newly-created words, such as “maimeng” (act cutely) and “pindie” (competition of family background).

No matter what type a popular Internet word is, its semantics in the context of a microblog is quite clear. Therefore, they should be involved in the 5W model. In other words, it should be classified into an 5W attribute. Some research work concentrates in particular on hot-word detection[29]. Therefore, we built a popular Internet word lexicon and integrated it into the 5WTAG algorithm to help the word segmenter recognize these more efficiently and accurately.

4.2.2 Popular Internet sentences
Popular Internet sentences are another interesting problem. The completeness in syntax and independence in semantics are the most two significant features of popular Internet sentences. Let us take “Yuanfang, ni zenme kan? (What is your opinion? Yuanfang)” as an example. It is an actor’s line from a famous Chinese TV series. When using this sentence, the microblogger just wants to express his or her concern about some event or ask for opinions from others. We can conclude that “Yuanfang” (a character’s name in the TV series) has nothing to do with the ongoing event. In most cases, popular Internet sentences should be pruned to avoid their influence on the extraction of the five Ws.

The number of newly-emerging popular sentences is far less than that of popular words, which makes it possible to collect them manually.

4.3 Segmenting microblogs
To construct candidate topic hashtags for a given microblog, we needed to segment the microblog into clauses. Each clause had relatively complete semantics and contributed one candidate hashtag. We had two categories of punctuation worthy of consideration as separators: termination punctuations, namely period, question mark, exclamation mark, and semicolon; and pause punctuations, including commas and all the termination punctuations.

Next, we discuss the influence of these different series of punctuations on microblog segmentation.

4.3.1 Using termination punctuations
According to the principles of Chinese grammar, clauses separated by termination punctuations are more complete in syntax. Therefore, candidate topic hashtags, which are extracted from termination-punctuation-based segmented clauses, may be more complete in semantics. However, this strategy is vulnerable to punctuation fault tolerance. Many microbloggers do not observe the principles of Chinese grammar in writing. Some microblogs are even separated by commas from the very beginning to the end. Such scenarios may result in semantic confusion if we segment them with termination punctuations.

4.3.2 Using pause punctuations
Comparatively speaking, it is more adaptive and avoids semantic confusion if microblogs are segmented with pause punctuations. However, the segmented clauses may be fragmentary in syntax, which will inevitably
weakens the semantical completeness of the topic hashtags.

4.4 Constructing candidate hashtags

We now constructed the candidate topic hashtags based on the segmented clauses. As is defined before, \( \langle w, \text{null} \rangle \) is a microblog without topic hashtags. After being segmented, it can be denoted as \( \langle \langle s_1, s_2, \ldots, s_l \rangle, \text{null} \rangle \), where \( \langle s_1, s_2, \ldots, s_l \rangle \) is the sequence of segmented clauses in their order of appearance. Let \( d \) denote a 5W word in \( \langle w, \text{null} \rangle \). It must be contained in a clause, say \( s_j \). Then, the word \( d \) can be extended and denoted as \( \langle d, \Psi_X(d), \text{ssid}, \text{loc} \rangle \). \( \Psi_X(d) \) is one of the five Ws that \( d \) is mapped to. ssid is the sequence number of the clause \( s_j \). loc is the position index of \( d \) in \( w \).

To guarantee semantic correctness as much as possible, we observed the following two rules when constructing the candidate topic hashtags.

Rule 1 Words that compose the same candidate topic hashtag should come from the same source clause.

Rule 2 Words in a candidate hashtag should keep their original order, as in the source clause.

We use \( h^* \) to denote a candidate topic hashtag. According to Rules 1 and 2, \( h^* \) is as follows:

\[
\begin{align*}
\langle d_1d_2\ldots d_m \rangle, & \quad \text{s.t. } \forall d_i, \forall d_j, i \neq j \nonumber \\
\{ & d_i.\text{ ssid} = d_j.\text{ ssid}, \nonumber \\
\{ & d_i < d_j \iff d_i.\text{ loc} < d_j.\text{ loc} \}
\end{align*}
\]

Here \( d_i < d_j \) means that the word \( d_i \) goes before \( d_j \) in microblogs. Note that since the candidate hashtags are derived from the segmented clauses, we may get more than one candidate hashtags. We need to pick out the most likely one for topic hashtag clustering.

5 Topic Hashtag Recommendation

In this section, we evaluate each of the obtained candidate hashtags on the basis of several parameters, namely semantic completeness and correctness, probability of occurrence of the hashtag mode, content-based importance, and location-based importance. We hope that the characteristics of machine-annotated topic hashtags agree with those of human-annotated ones as much as possible. To this end, we needed to first understand the statistical distribution of human-annotated hashtags.

5.1 Distribution of artificial hashtags

As mentioned in Section 1, only 23% of tweets have at least one topic “#hashtag”. In Sina Weibo, this ratio drops to 12.9%. We examined over 11 000 distinct human-annotated hashtags and achieved the length distribution as well as the presentation ratio of every attribute in the five Ws.

5.1.1 Length distribution of artificial hashtags

The length of a hashtag is defined as the number of Chinese characters in it. We analyzed 11 008 distinct artificial hashtags from Sina Weibo, and achieved the length distribution in Table 4.

Table 4 shows that most artificial hashtags contained 1 to 14 Chinese characters, which accounts for more than 99% of the total. The mean value and variance of these hashtags are \( \mu = 5.33 \) and \( \sigma = 4.57 \), respectively.

5.1.2 Presentation ratio of the attributes in five Ws

According to our observation regarding human-annotated hashtags, words with different attributes in the five Ws were not present at the same probability in topic hashtags. This inspired us to examine the presentation ratio of every attribute in the five Ws.

Definition 6 The Presentation Ratio of an attribute is the frequency at which the words of this attribute are selected for the human-annotated topic hashtags.

Assume that there are a group of microblogs \( \Gamma \).

\[
\sum_{\forall (w_X, h_X) \in \Gamma} \sum_{\forall d_i \in \Psi_X(w_X)} (d_i \in h_X) \times \frac{\text{the times of that the words in } \Gamma \text{ belonging to the attribute } X \text{ are selected for the human-annotated topic hashtags } h_x.}{\sum_{\forall (w_X, h_X) \in \Gamma} \Psi_X(w_X)}
\]

The presentation ratio of the attribute \( X \), denoted as \( p(X) \), is the ratio of these two values, i.e.,

\[
p(X) = \frac{\sum_{\forall (w_X, h_X) \in \Gamma} \sum_{\forall d_i \in \Psi_X(w_X)} (d_i \in h_X)}{\sum_{\forall (w_X, h_X) \in \Gamma} \Psi_X(w_X)}
\]

Obviously, the higher this ratio, the more frequently the words with the attribute \( X \) are present in hashtags.

To obtain statistical results of \( p(X) \), we measured

| Number of characters | Ratio (%) | Number of characters | Ratio (%) |
|---------------------|-----------|---------------------|-----------|
| 1                   | 0.7       | 8                   | 5.5       |
| 2                   | 8.9       | 9                   | 4.4       |
| 3                   | 15.0      | 10                  | 3.1       |
| 4                   | 26.1      | 11                  | 2.6       |
| 5                   | 12.4      | 12                  | 1.6       |
| 6                   | 9.8       | 13                  | 1.4       |
| 7                   | 7.9       | ≥14                 | 0.6       |
11,008 artificial hashtags and calculated the $p(X)$ of every attribute in 5W according to Eq. (2). Table 5 shows the results.

Table 5 reflects the talking habits of people quite clearly: when we describe an event, “Someone did something” is the most important information that we want to convey. So, the words “Who”, “hoW”, and “What” are more likely to be present in artificial hashtags than the other two attributes.

5.2 Recommendation calculation

In this section, we introduce metrics for evaluating the obtained candidate topic hashtags $h^*$, namely semantic completeness and correctness, probability of the candidate hashtag mode, content-based importance, and location-based importance.

5.2.1 Semantic completeness

Semantic completeness defines that to what extent a candidate topic hashtag can represent a whole event. To quantify the semantic completeness, we introduced the concept of the hashtag mode.

**Definition 7** The Hashtag Mode of a hashtag is the combination of the 5W attributes that the words in this hashtag correspond to.

We used $M(h^*)$ to denote the hashtag mode for $h^* = \langle d_1 d_2 ... d_m \rangle$. $M(h^*)$ is calculated with Eq. (3).

$$M(h^*) = \{\Psi_X(d_1), \Psi_X(d_2), ..., \Psi_X(d_m)\} \quad (3)$$

As the five Ws are requisite elements for event description, the more 5W attributes a candidate topic hashtag contains in its hashtag mode, the more details it conveys about the event. Therefore, we concluded that the semantic completeness of $h^*$ is directly proportional to cardinality of $M(h^*)$, i.e., $|M(h^*)|$.  

5.2.2 Semantic correctness

Semantic correctness indicates if a hashtag is semantically understandable. In natural languages, words that are logically relevant are closer to each other. This provides an effective approach to determine the semantic correctness of $h^*$. We introduced the Average Word Interval $\overline{\lambda}$ to $h^* = \langle d_1 d_2 ... d_m \rangle$. $\overline{\lambda}$ can be calculated by Eq. (4).

$$\overline{\lambda} = \frac{\sum_{i=1}^{m-1} (d_{i+1}.\text{loc} - d_i.\text{loc} - \text{length}(d_i))}{\text{length}(h^*)} \quad (4)$$

The semantic correctness of a candidate hashtag is inversely proportional to $\overline{\lambda}$. In other words, the more widely the words in $h^*$ scatters in its source clause, the worse semantic correctness $h^*$ may have.

5.2.3 Probability of hashtag mode

Better semantic completeness requires more 5W attributes in the hashtag mode as it means more details about the event. However, Table 5 shows the opposite phenomenon in human-annotated hashtags: Not all the attributes in 5W are necessarily present in artificial topic hashtags. According to our statistics, only around 1.2% of artificial hashtags contained more than four attributes in their hashtag mode. Furthermore, given the attribute presentation rate in Table 5, we calculated the probability of hashtag mode of $h^*$ as follows:

$$p(M(h^*)) = \prod_{X \in M(h^*)} p(X) \prod_{X \notin M(h^*)} (1 - p(X)) \quad (5)$$

$p(X)$ is the attribute presentation ratio of the attribute $X$. As the values of $p(X)$ are all less than 1, the shorter and simpler hashtags rank higher. This is exactly consistent with the tendency of simplification in artificial hashtags.

5.2.4 Content-based importance

When people describe an event or make a comment, it is common for them to repeat some key words for emphasis. Therefore, we concluded that the content-based importance of a candidate topic hashtag be directly proportional to its Keyword Density$^{[30]}$.

Let $\rho_{h^*}$ denote the content-based importance of the candidate hashtag $h^* = \langle d_1 d_2 ... d_m \rangle$, and $\rho_{d_i}$ denote Keyword Density of the word $d_i$ in $h^*$. Obviously, $\rho_{h^*}$ is the continued product of multiple $\rho_{d_i}$.

$$\rho_{h^*} = \prod_{\forall d_i \in h^*} \rho_{d_i} = \prod_{\forall d_i \in h^*} \frac{\text{count}(d_i | w)}{\sum_{\forall d \in w} \text{count}(d | w)} \quad (6)$$

Here, $\text{count}(x | y)$ is the number of occurrences of the word $x$ in the text $y$.

5.2.5 Location-based importance

Due to the limitation on characters, microblogs and short-message-style news are usually structured in the format of an “inverted pyramid”, i.e., the most important information is placed first within a text, and information of decreasing importance is placed in subsequent paragraphs. So, we can evaluate the
importance of the candidate topic hashtag $h^*$ according to  
its source clause’s location within the microblog.

For $h^* = \langle d_1d_2...d_m \rangle$, its location, denoted as $h^*.\text{loc}$, is defined as being the same as that of its source clause in a microblog, i.e.,

$$h^*.\text{loc} = d_i.\text{ssid}, \forall d_i \in h^* \quad (7)$$

We studied five metrics that influence the recommendation of the candidate hashtag $h^*$. In summary: First, the semantic completeness and correctness are semantics-oriented evaluation parameters. Second, the probability of a hashtag mode is an empirical-data-oriented evaluation parameter. Third, the content-based and location-based importance are syntax-oriented evaluation parameters. We present a comprehensive recommendation function for evaluating the candidate topic hashtag $h^*$ as Eq. (8).

$$S(h^*) = |M(h^*)|\cdot (1-\bar{\lambda})\cdot p(M(h^*))\cdot \rho_h \cdot \frac{1}{h^*.\text{loc}} \quad (8)$$

The candidate hashtag with the highest $S(h^*)$ will be recommended as the topic hashtag for $w$.

5.3 Compressing overlength hashtags

Due to the positive correlation between the completeness of a candidate hashtag and its recommendation, the longer candidates may defeat the shorter ones. However, long candidate hashtags are not advantageous in the classification and retrieval of microblogs. Moreover, they are not in accord with the short and simple characteristics of human-annotated hashtags. We set a threshold $T$ to limit the length of automatically-annotated hashtags. In this paper, we make use of the statistical results in Table 4, and set $T$ as $[\mu + \sigma] = 10$. Next, we introduce two strategies for compressing overlength hashtags.

**Method 1 Length-based Compression.** The basic idea of this method is to remove the longest words from $h^*$ iteratively until it meets the requirement of length constraint. The advantage of this method is the simplicity in implementation and efficiency in execution; the weakness is that it neglects the semantics of words and may decrease the semantic correctness of $h^*$.

**Method 2 Semantics-based Compression.** The basic idea of this method is to remove all the words of one attribute in 5W once to shorten the length of $h^*$. Considering the statistical result in Table 5, we should remove the words of “Where” first, then the others successively in the order of attribute presentation rate. Compared with the former method, the latter takes the semantics of words into consideration, but it may weaken the completeness of $h^*$, and lead to the failure of $h^*$ in event description.

5.4 Candidate hashtag clustering

So far, each microblog has been assigned with a recommended topic hashtag from the algorithm 5WTAG. This leads to a new challenge, too many different hashtags. Even microblogs on the same topic may have different hashtags for their diverse content. We needed to cluster the recommended hashtags into coherent groups. We used Affinity Propagation as our clustering algorithm as it does not require a prespecified number of clusters. We modeled the dataset of all recommended candidate hashtags as $\langle \mathcal{H}, M \rangle$, where $\mathcal{H}$ is the candidate hashtag set, and $M$ is hashtag pairwise relatedness matrix. Affinity Propagation takes as input a collection of real-valued similarities between pairs of hashtags, i.e., $M$. The output is a good set of exemplars and corresponding clusters.

As for hashtag relatedness, we had several metrics to measure the similarity between two hashtags such as Jaccard distance, Levenshtein distance, and cosine distance. However, they all lack consideration of the semantics of the 5W words contained in the hashtags. Consider two people asked to describe the same event, although their descriptions may differ in content, they must be the same in the core details of the event, i.e., the words of the 5W attributes. So, we can measure the relatedness of two hashtags with the Hashtag Mode as follows:

$$\text{Sim}(h_i^*, h_j^*) = \frac{|M(h_i^* \cap h_j^*)|}{|M(h_i^* \cup h_j^*)|} \quad (9)$$

where $|M(h_i^* \cap h_j^*)|$ is the number of the 5W attributes which the shared words by $h_i^*$ and $h_j^*$ correspond to, and $|M(h_i^* \cup h_j^*)|$ is the total number of the 5W attributes which are involved in $h_i^*$ or $h_j^*$.

6 Experiment Evaluation

6.1 Experimental dataset and methodology

In this section, we present the evaluation results of the real microblog dataset collected from Sina Weibo. Sina Weibo is the largest microblogging service provider in China, with 56.5% of market share based on active users, and 86.6% based on browsing time[32]. Table 6 shows details of the experimental data.

In our experiments, two methods were implemented: 5WTAG-P (Period) and 5WTAG-C (Comma). The former uses termination punctuations to segment the
microblogs, while the latter uses pause punctuations. We regarded the dataset of artificial hashtags as our baseline and verified the 5WTAG algorithm in four aspects: semantic correctness, quality of recommendation, performance of two compression strategies, and accuracy of topic-oriented hashtag clustering.

6.2 Experimental results

6.2.1 Semantic correctness evaluation

The semantic correctness of a candidate topic hashtag includes two aspects: (1) whether it is semantically understandable to users, and (2) whether it is the accurate abstract of its source clause. Theoretically, the semantic correctness evaluation can only be conducted manually. However, to promote efficiency, we partially adopted automatic evaluation. For that reason, we proposed two principles:

**Principle 1** If a candidate hashtag contains only one 5W word, it is semantically correct.

**Principle 2** If the average word interval of a candidate hashtag is 0, it is semantically correct.

We randomly sampled 1000 microblogs from our dataset and constructed candidate hashtags for each using the methods 5WTAG-P and 5WTAG-C. Table 7 shows the evaluation results.

Table 7 shows that 5WTAG-C performs better than 5WTAG-P for semantic correctness. As discussed in Section 4, 5WTAG-C segments a microblog into finer-grained clauses, which resulted in more candidate topic hashtags than 5WTAG-P. 5WTAG-C can effectively resolve long sentences with multiple semantics, and thus it avoids semantics confusion better than 5WTAG-P. The user’s arbitrary and ill-formed use of punctuation in microblogs is the predominant reason that makes 5WTAG-C more advantageous for semantic correctness. However, more candidate hashtags possibly mean worse semantic completeness.

To validate our prediction, we further tested the cardinality of the hashtag mode, the length of 5WTAG-C and 5WTAG-P, and the human-annotated hashtags. Figure 3 shows the results.

From Fig. 3a, we can see that most of the artificial hashtags are relatively short. This is because humans have the exclusive ability to associate and summarize. 5WTAG-P is inclined to create longer hashtags because of the larger-grained microblog segmentation with termination punctuations. Figure 3b shows the semantic completeness of artificial, 5WTAG-C, and 5WTAG-P hashtags. The X-axis is the number of attributes contained in candidate topic hashtags, i.e., the cardinality of the hashtag mode. Figure 3b can exactly explain the phenomenon in Fig. 3a: human-annotated hashtags are relatively shorter and simpler in most cases. Most microbloggers are more willing to use one or two keywords to represent the whole event. Besides, due to the different granularity of microblog segmentation, 5WTAG-C creates more hashtags with a single attribute, while 5WTAG-P creates more combination hashtags with multiple attributes.

| Method       | Number of candidate hashtags | Semantical correctness (%) |
|--------------|------------------------------|----------------------------|
| 5WTAG-P      | 2566                         | 42.1                       |
| 5WTAG-C      | 4218                         | 69.7                       |

Fig. 3 Comparison on artificial, 5WTAG-P, and 5WTAG-C hashtags.
6.2.2 Recommendation quality evaluation

We conducted the recommendation quality evaluation based on 500 pieces of microblogs sampled randomly from the dataset. For each of the microblogs, we used 5WTAG-C and 5WTAG-P to create candidate topic hashtags, from which we manually chose the most satisfying candidate as its topic. We then ran the recommendation function to assign a recommendation score to each of the candidate topic hashtags. Next, we examined if the human-chosen hashtag was the same as the candidate hashtag that ranked highest according to the recommendation function. If they were the same, recommendation function was considered accurate. Table 8 shows the results.

Table 8 shows that 5WTAG-P performs better than 5WTAG-C for recommendation quality. The reason is that the candidate hashtags created by 5WTAG-C are often too simple to describe the whole event. When 5WTAG-C segments microblogs in a finer-grained way, it breaks the overall semantics of the microblog at the same time. Topic-related words are scattered into more clauses, which means that every single candidate hashtag may lose some key details about the event. So, it leads to more failure in recommendation accuracy. According to the experiments above, we can conclude that 5WTAG-C performs better for the semantic correctness, while 5WTAG-P is better for semantic completeness and recommendation accuracy.

6.2.3 Compression method evaluation

In the third experiment, we tested the performance of two methods for compressing overlength hashtags. We manually chose 500 hashtags, generated by 5WTAG-P, to guarantee their semantic correctness. Each of these hashtags contained more than ten Chinese characters, and their topics were uniformly distributed across multiple social fields, namely sports, politics, economy, entertainment, culture, and military. Then we ran the two compressing methods and examined whether a compressed hashtag still conveyed the main idea of the microblog. If so, it was still semantically correct. Table 9 shows the results.

Table 9 shows that the semantic correctness of the length-based compression method drops dramatically. This is because this method totally ignores the semantics of words. When the words vital to semantics presentation are removed, the whole topic hashtag may be no longer understandable. Among semantics-based compression methods, the semantic correctness of “What”-based pruning method drops the most. The reason is that “What”-related words are the most necessary elements in event description. The result in this experiment is consistent with the statistics in Table 5. The importance of “What”-related words gives more opportunities for them to be included in topic hashtags. The experiment result in Table 9 suggests that if we have to shorten a candidate topic hashtag, we should adopt a hybrid compression method: pruning the longest words successively in the attribute order of “Where”, “When”, “How”, “What”, and “Who” until the candidate hashtag meets the length constraint.

6.2.4 Topic clustering evaluation

Here we present the evaluation result of topic clustering based on the recommended hashtags. Firstly, we needed to build a standard dataset as a ground-truth for comparison. We manually select 10 hot topics from the 11008 distinct artificial hashtags such as “#Zhongshenzhinu#” (Wrath of the Titans), “#Lixingaiguo#” (Rational patriotism), “#Beijingkongqiwuranzhishu#” (Beijing background air pollution index), and “#Shejianshangdechunjie#” (a bite of Chinese New Year). A total of 1473 microblogs resulted and we ran 5WTAG-P to extract the topic hashtag for them. Then, we calculated the similarities between the pairs of obtained recommended hashtags and generated the relatedness matrix. Finally, we ran Affinity Propagation to cluster these hashtags and compared the clustering results with our standard dataset. The result is in Table 10.

From Table 10 we can see that the Jaccard distance, Levenshtein distance, and cosine distance methods produced far more clusters than the standard dataset. This indicates that content-based similarities are not an effective measure for the relatedness between the recommended hashtags. Accordingly,
they cannot achieve high accuracy in topic hashtag clustering. The Hashtag Mode approach emphasizes the semantic similarity. For example, “Shenyang” and “LiaoningShenyang” are different in terms of content, but are exactly the same in the semantics of the attribute “Where” in 5W. Therefore, the Hashtag Mode approach dramatically improves the clustering performance of Affinity Propagation.

7 Conclusion and Future Work

In this paper, we introduce a novel algorithm, named 5WTAG, for determining topic hashtags in Chinese microblogs. We modeled Chinese microblogs using the 5W model. The concept of five Ws comes from journalism and represents five questions, the answers to which are considered to be the basic elements for event description. We reveal the justification for modeling microblogs with the 5W model by comparing the similarity between them and short-message-style news. We introduce the detailed procedure of the algorithm 5WTAG, including segmenting a microblog into clauses and constructing candidate topic hashtags. Next, we put forward a quantitative method for recommendation computation, which integrates semantics, syntax, and empirical data. We propose using the Affinity Propagation algorithm to cluster the recommended hashtags and suggest a novel 5W-based approach to measure the topic similarities of pairs of hashtags.

We are developing several applications based on microblogs with 5W topic hashtags. For example, when a public event happens, many witnesses on the spot may release microblogs describing the same event. These microblogs are possibly written differently and may have added individual comments, but they are quite similar in the five Ws. Therefore, we can merge them and detect public events at the first instance. In addition, this can be applied to the identification of Internet rumors. All these applications are in the plan of our near future work.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 61173027) and the Northeastern University Fundamental Research Funds for the Central Universities (Nos. N150404012 and N140404006).

References

[1] A. Bruns and J. Burgess, The use of twitter hashtags in the formation of ad hoc publics, in Proc. 6th European Consortium for Political Research General Conf., Reykjavik, Iceland, 2011, pp. 1–9.
[2] Twitter Help Center, http://support.twitter.com/articles/49309, what_are_hashtagsymbols, 2016.
[3] X. F. Meng, F. R. Wei, X. H. Liu, M. Zhou, S. J. Li and H. F. Wang, Entity-centric topic-oriented opinion summarization in twitter, in Proc. 18th Annual ACM Conf. on Knowledge Discovery and Data Mining, New York, NY, USA, 2012, pp. 379–387.
[4] I. Yap, H. T. Loh, L. X. Shen, and Y. Liu, Topic detection using MFSS, in Proc. 19th Int. conf. on Advances in Applied Artificial Intelligence, Berlin, Germany, 2006, pp. 342–352.
[5] Y. W. Seo and K. Sycara. (2004, Jan.), Text clustering for topic detection. Available: https://www.researchgate.net/publication/2901255_Text_Clusterin_for_Topic_Detection.
[6] X. Y. Zhang and T. Wang, Research of technologies on topic detection and tracking, Journal of Frontiers of Computer Science and Technology, vol. 3, no. 4, pp. 347–357, 2009.
[7] T. Hofmann, Probabilistic latent semantic indexing, in Proc. 22nd Int. ACM SIGIR Conf., Berkeley, CA, USA, 1999, pp. 50–57.
[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research, vol. 2003, no. 3, pp. 993–1022, 2003.
[9] G. Xu and H. F. Wang, The development of topic models in natural language processing, (in Chinese), Chinese Journal of Computers, vol. 34, no. 8, pp. 1423–1436, 2011.
[10] W. X. Zhao, J. Jiang, J. S. Weng, J. He, E. P. Lim, H. F. Yan, and X. M. Li, Comparing twitter and traditional media using topic models, in Proc. 33rd European Conf. on Advances in Information Retrieval, Dublin, Ireland, 2011, pp. 338–349.
[11] J. S. Weng, E. P. Lim, J. Jiang, and Q. He, Twitterrank: Finding topic-sensitive influential twitterers, in Proc. 3rd Int. Conf. on Web Search and Web Data Mining, New York, NY, USA, 2010, pp. 261–270.
[12] X. H. Yan, J. F. Guo, Y. Y. Lan, and X. Q. Cheng, A bitemr topic model for short texts, in Proc. 22nd Int. Conf. on World Wide Web, Rio de Janeiro, Brazil, 2013, pp. 1445–1456.
[13] X. Q. Cheng, X. H. Yan, Y. Y. Lan, and J. F. Guo, BTM: Topic modeling over short texts, IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 12, pp. 2928–2941, 2014.
[14] R. Mehrotra, S. Sanner, W. Buntine, and L. X. Xie, Improving LDA topic models for microblogs via tweet pooling and automatic labeling, in Proc. Int. ACM SIGIR
Zhibin Zhao received the PhD degree from Northeastern University, China, in 2007. He is currently an associate professor with College of Computer Science and Engineering, Northeastern University. He was a visiting scholar at Computer Science Department, Illinois Institute of Technology from 2012 to 2013. His research interests include distributed computing, Web data mining, and big data management.

Jiahong Sun received the BS degree from Shenyang Normal University in 2014. She is now a master student in College of Computer Science and Engineering, Northeastern University. Her research interests are Web systems and Web data mining.

Lan Yao received the PhD degree from Northeastern University, China, in 2008. She is currently an associate professor with College of Computer Science and Engineering, Northeastern University. She was a visiting scholar at Computer Science Department, Illinois Institute of Technology from 2011 to 2012. Her research interests are Web data integration, cloud computing, and wireless sensor networks.

Xun Wang is currently a master student in College of Computer Science and Engineering, Northeastern University. He received the bachelor degree from Hebei GEO University in 2014. His research interests are Web data mining and deep learning.
Jiahong Chu is currently a master student in College of Computer Science and Engineering, Northeastern University. She received the bachelor degree from Inner Mongolia University in 2013. Her research interests are Web data management and information system development.

Huan Liu is currently a master student in College of Computer Science and Engineering, Northeastern University. He received the bachelor degree from North China University of Water Resources and Electric Power in 2014. His research interests are Web data integration and cloud computing.

Ge Yu received the PhD degree in computer science from Kyushu University of Japan in 1996. He is currently a professor with Northeastern University of China. He serves on the editorial board of IEEE Transactions on Knowledge and Data Engineering, and has served on the program committee for many international conferences (including VLDB, ICDE, CIKM, ER, DASFAA, ADC, etc). He has published more than 200 papers in refereed journals and conferences. He is a member of the IEEE Computer Society, IEEE, ACM, and CCF. His research interests include distributed and parallel database, OLAP and data warehousing, data integration, graph data management, etc.