Overview of Topic Detection and Tracking of Methods for Microblogs

Song WANG¹,a and Xue-guang ZHOU²,b,*

¹School of Economics and Management, Chuxiong Normal University, Chuxiong, Yunnan, China
²Department of Information Security, Naval University of Engineering, Wuhan, Hubei, China

a36606469@qq.com, bzxg196610@hotmail.com
*Corresponding author

Keywords: Microblog, Online social networks, Topic detection and tracking.

Abstract. There are still relatively few research studies on topic detection and tracking of microblogs. Traditional methods applied to microblogs cost too much calculation and the rate of accuracy is low. Therefore, combined with characteristics of large data of documents, short single documents, few hot topics and colloquial contents, the article studies and fulfills methods of topic detection and tracking of microblogs.

Introduction

The concept of Topic Detection and Tracking (TDT) firstly originated in 1996 when the United States Defense Advanced Research Projects Agency (DARPA) proposed that in the light of their demands they should develop new technology which can automatically identity topics of streams of news data without any manual intervention to assist people to face increasingly serious information explosion of internet, automatically identify new topics of information flow of news media and continually track known topics [1-2].

Since microblog was born in 2006, a variety of platforms have appeared and users have been increasing continually. Examples of online social networks are Facebook, Twitter, Sina and Tencent. The research progress was analyzed from five aspects including general theory, emergency, public opinion, and the relationships of microblog and individual and social organization [3-7].

TDT and microblog can now achieve more intelligent integration and managements through Natural Language Processing (NLP) technology. The emphasis of this article is to summarize the calculation of similarity of concepts of characteristics of microblog and accurately judge the relevance of topics by TDT technologies.

The Calculations of Topic Detection and Tracking

The Calculation of Topic Detection

The research of TDT is divided into five parts, which include story segmentation, relevance detection of stories, topic detection, topic tracking and cross-language technology aimed at these tasks. In the five parts of the study of TDT, relevance detection of stories is the most important study project above all. Only through enhancing the accuracy of judgment of similarity of reports can we efficiently detect and track follow-up topics. In view of people become thirsty for some tool hat which can conveniently and accurately gather information of related topics in order that information is accessible. The technology of TDT emerges as a result. TDT originates from the former Event Detection and Tracking (EDT) which can offer people the following tracking of known topics and automatic detection of new topics of various information flow of news media such as television, broadcast, websites, etc[8].

Indifference of EDT, objects detected and tracked by TDT can be expanded to topics with more
extended relevance, theory study and application corresponding to which has transformed from
traditional detection of events to topic detection and tracking of emergency and follow-up reports.
Topic detection and topic tracking are two subtasks of TDT with extremely broad prospect in fields
of financial securities, information security and industry survey.

The main task of relevance detection is to detect that whether two randomly chosen reports
explain the same topic. Traditionally in TDT study based on probability and statistics, relevance
between reports and topics or reports and reports are judged by checking the coverage ratio of
common feature sof them. Most studies on relevance detection focus on text representation and
feature selection. James Allan is one of the first scholars solving TDT problems with NLP, who
adopted VSM technique to represent topics and reports and gave name entity higher weight to
perform the task of NED in TDT technique[1]. Nallapati improved this method that he firstly
distributed features to different syntax categories, such as noun classes and verb classes in part of
speech, as well as time, site and name classes in name entity[9].

On this basis he adopted the method of probability statistics of syntax model to estimate
probability whose features originated from different syntax categories and label the weight of
features. Juha Makkonen proposed to construct semantic category in the solution of problems of
relevance detection in TDT by judging similarity of documents through similarity of semantic
category[10].

As semantic relation is the relation between meanings of words, which can be presented by
collections of synonyms, semantic relations is regarded as guide of collections of synonyms. As
semantic relation includes essynonymy, antonymy, hyponymy and meronomy, it shows from
features of WordNet that WordNet is extraordinarily similar to ontology[11].

TDT is the study against natural language reports, and in the aspect of similarity detection of
texts, it can expand semantic features with geographic ontology for the similarity of geographic
noun sand calculate its information of concepts through WordNet for concept sof name entity and
behavior in reports[12].

The Calculation of Topic Tracking

The fundamental design idea of the system of topic tracking is that it firstly provides one group of
training news reports, after preprocessing texts, it presents news texts into the known topic model.
With the same process of streams of news reports in text corpus, it gains models of unknown topics.
Afterwards, it calculates the similarity of the two models and judges whether this news report
belongs to the known topic comparing modules through threshold value.

In the text preprocessing module of the system of topic detection and tracking, processing word
segmentation and annotation of the part of speech of news report texts. In recent years, a series of
evaluation by TDT organization has used specified news corpus provided by TDT organization. In
this corpus, each piece of news text has been processed by word segmentation and annotation of the
part of speech, and each text possesses a standard format. Above all one document picks up feature
words by word segmentation and annotation of the part of speech. It is important to note that
ambiguity is common in Chinese texts, the segmentation of Chinese sentences is difficult to some
degree, and there are a number of studies in the field of disambiguation in china. After word
segmentation of documents, the next step is to identify the part of speech of word groups, which
need search semantic dictionary to give relevant word groups the part of speech. Moreover, it has to
distinguish part of speech by the sentence structure to several part sof speech of word groups to find
the part of speech according with syntactic logical relation, where a system of syntactic
identification. In the process of syntactic discriminant, it should abandon ords whose contributions
to text feature can be neglected and reserve relatively useful feature words for documents such as
word groups describing time and geographic locations, noun word groups, names, terms and verb
word groups. These words will be classified according to semantic types previously classified and calculated its contribution to the text.

For word groups describe periods, they should be normalized in preprocessing stage as much as possible, which is to map the min the global unified timeline. If there is certain definite time in the texts which can map certain time description in the unified timeline, the time description should be remained or abandoned as uncertain time description cannot be calculated its similarity.

The system model of TDT is mainly composed of probability model and vector space model. Vector space model is constantly developed, where several vectors can present a text and then the technique that presents texts with semantic categories appears. In the study of Juha Makkonen[10], one document is divided into four semantic categories: LOCATIONS, TEMPORALS, NAMES, TERMS. The division of the four semantic categories is concluded after the observation of facts that certain event is related to certain period of time, site, specific person or thing. The division after the management of the feature is still not that meticulous. Through analysis the event is not only related to certain specific person or thing as well as some specific behavior. One event is caused by certain behavior in certain time and space, which involves four features, that is time, site, person or thing related to the event and behavior generated. In the analysis, the four features are indispensable and the only means to decide an event, and the four features are necessary elements for an event. Therefore, it is a proper method of division to classify the four features as semantic category. Therefore we consider the feature words as LOCATIONS, TEMPORALS, NAMES and TERMS.

Another advantage of the division of semantic category is that in the analysis of sentence structure, it will involve subject and predicate which can be seen below. The composition of sentence triples require noun and verb feature words, which are in ENTITIES and ACTIVITIES categories[13].

After preprocessing module of texts, texts with named entities are received. In 1988, Salton and Buckley gave a new approach to computer the text retrieval automatically [14].

The kernel of the approach is as below:

\[ w_{ik} = TF \times IDF = tf_{ik} \times \log \left( \frac{N}{n_k} + 0.01 \right) \]  

(1)

In the process of realizing the algorithm, the text feature selection method based on document frequency is TF-IDF weighting strategy. TF is the term frequency of occurrence of words in the current text representing the importance of words in the description of the current text. IDF is the inverse document frequency of text figure with words to weaken the importance of frequent words in corpus as frequent words are regarded as those lacking distinguishing ability.

After the weight of all words is received, there are two methods to constitute space vectors with words.

(1) After removing function words and stop words, it will sort the remaining words according to the weight of each word and select the first N words of the highest weight to constitute space vectors to represent the current texts of news report.

(2) All words are taken as features to constitute vectors. Finally each piece of news report is represented as the vector like \[ V_{d_i} = \left\{(T_{1d}, W_{1d}), (T_{2d}, W_{2d}), \ldots, (T_{id}, W_{id}), \ldots, (T_{nd}, W_{nd})\right\} \]  

(2)

After receiving the model of topics and reports, it will calculate the similarity between them. When calculating the similarity between texts, the actual quantities of words in the model prevail
because of different lengths of news reports. Here feature quantitative value is assumed character_num and the quantities of feature words of current topics are N.

The Calculations of Topic Detection and Tracking for Microblogs

The Characteristics of Topics for Microblogs

Information filtering of microblog and broad user group of microblog attract a number of hackers to launch spam on the platform of microblog. Recently each platform adds blocking functionality, but the effect is not ideal that plenty of spam still exists on microblog. On the foundation of sentiment analysis of microblog, users express their feelings from their hearts on microblog. It can provide guidance for product marketing and public opinion monitoring after the analysis of sentiment analysis of microblog, which studies topic detection and tracking method combined with features of microblog users and structures. Up till now there are studies on topological structure of microblog and users, influence of users shared by scholars and the combination of traditional topic detection and structure and users of microblog application can be considered to conclude a better method of topic detection and tracking of microblog.

Hierarchical-Clustering Algorithm for Microblog Topic Detection

The topic detection for microblog belongs to off-line backtracking detection, which focuses on newly increased microblogs in one period of time as one subtask in the study system of topic detection and tracking for microblogs. Traditional hierarchical-clustering algorithm is divided into filtration hierarchical clustering and split hierarchical clustering. The topic detection based on hierarchical clustering in the article belongs to backtracking detection mainly in advantage of the idea of filtration hierarchical clustering which belongs to the strategy from the bottom up. Initially it sets each object as a topic cluster and combines two most similar clusters to form one high-level topic cluster until the highest level of the topic cluster reaches the default which meet terminal condition. Methods of hierarchical clustering in the majority of topic detection belong to the category and they only differentiate in the definition of similarity between topic clusters, seen in Figure 1.
Hierarchical clustering of microblogs in traditional algorithm of topic tracking is as follows.

1. Preprocessing: Preprocess microblogs, select feature items and represent them as vector space model (VSM).

2. Initialization: calculate the similarity of all news texts in microblog set and build similarity matrix between microblog set.

3. Selection: search two most similar pieces of microblog in the matrix and combines them to form a topic category.

4. Updating the matrix: set combined topic category as a new object in the set and recalculate the similarity to form a new similarity matrix in microblog set.

5. Repeat: If the quantity in the similarity matrix is more than 1 or doesn’t reach the specified terminal condition, turn to (3).

6. Termination: Finally a hierarchical clustering tree forms.

**SVM Algorithm for Microblog Topic Tracking**

Literature [15] used VSM for representing topics, and then developed topic tracking system by use of support vector machines (SVM) algorithm for text classification. This system included three parts, one was topics representation model, the second was SVM algorithm for text classification, and the third was TDT evaluation. The experimental results and performance analysis showed the algorithm was good at the topic tracking.

Zhou proposed a new method based on relationship analysis using dependent sentence pattern for sub-Topic Detection and Tracking (sTDT) in [16], which constructs feature dimensions to generate the global vectors according to the increment of TF-IDF, and then creates the partial adjoin map based on the connection weights within the time window and decreases the dimensions through dependent sentence pattern. Experiments show the proposed method transferred the text from linear to plane structure, and extracted the subtopics effective. Considering that sTDT are similar to microblogs, this method can also be used for microblog TDT.

The identification of emotional tendencies for microblog has important practical significance. Li proposed a method of sentiment analysis for microblog based on SVM in [17], used various features including word, speech, emotional word, negative word, adverb of degree and special symbols. They designed different combinations of features to make the effect optimal through multiple sets of experiments. The accuracy of SVM reached 88.72% using the combination of speech, sentiment word and negative words.

To solve the Chinese microblogs sentiment classifications automatically, an unsupervised topic sentiment model, UTSM, is proposed based on dependency parsing [18]. In this paper, different feature combinations are used to complete the part of speech feature comparison experiment. Some results of the experiment are shown in Table 1.

| No. | Feature combination | Precision for positive corpus | Precision for negative corpus | Precision recall | F-measure |
|-----|---------------------|-------------------------------|-------------------------------|-----------------|-----------|
| 1   | SVM[17] Part of speech +emotional words +negative words | 87.95 | 89.59 | 88.72 | / | / |
| 2   | Part of speech +emotional words +negative words +adverb of degree | 87.84 | 88.87 | 88.32 | / | / |
From Table 1, it can be seen that the accuracy of SVM is slightly better than UTSM, but the biggest advantage of the latter is that it can automatically process massive microblogs. Therefore, this kind of algorithm can be considered to be introduced into microblog topic tracking processing.

Summary

The study of microblog topic detection and tracking has to be promoted by natural semantic understanding technique field, which will be seen that in future the technique will be applied widely in the study of TDT. On the basis of the study, it studies the algorithm of topic detection and tracking combined with methods of text classification to render the result of topic detection and tracking of text topics currently analyzed in different categories. At last different algorithms of microblog topic detection and tracking algorithms are discussed.

Acknowledgement

The authors would like to thank the anonymous reviewers for their detailed reviews and constructive comments, which have helped improve the quality of this paper. This work has been supported by postgraduate project of military science of PLA under Grant No. JY2018C148.

References

[1] James Allan, Jaime Carbonell, George Doddington, et al. “Topic detection and tracking pilot study: final report” [C]. In Proc. of the DARPA Broadcast News Transcription and Understanding Workshop. Virginia: Lansdowne, February 1998, San Francisco: Morgan Kanfmann Publishers, 1998: 194-218.

[2] James Allan. “Topic Detection and Tracking: Event-based Information Organization”[M]. Berlin/Heidelberg: Springer Science & Business Media, 2012.

[3] Efron M. “Information Search and Retrieval in Microblogs”[J]. Journal of the American Society for Information Science and Technology, 2011, 62(6): 996-1008.

[4] Lee R, Wakamiya S, Sumiya K. “Discovery of Unusual Regional Social Activities Using geo-tagged Microblogs”[J]. World Wide Web, 2011, 14(4): 321-349.

[5] Takehara T, Miki S, Nitta N, et al. “Extracting Context Information from Microblog Based on Analysis of Online Reviews” [C]. 2012 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). IEEE, 2012: 248-253

[6] Stieglitz S, Dang-Xuan L. “Emotions and Information Diffusion in Social Media-Sentiment of Microblogs and Sharing Behavior” [J]. Journal of Management Information Systems, 2013, 29 (4): 217-248.

[7] Oulasvirta A, Lehtonen E, Kurvinen E, et al. “Making the Ordinary Visible in Microblogs” [J]. Personal and ubiquitous computing, 2010, 14(3): 237-249.
[8] Allan James, Papka R, Lavrenko V. “On-line new event detection and tracking”[C]. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1998:37-45.

[9] Ramesh Nallapati, Ao Feng, Fuchun Peng, and James Allan. “Event threading within news topics”[C]. In CIKM, 2004, pp: 446-453.

[10] Juha Makkonen. “Investigations on event evolution in TDT”[C]. In Proceedings of Student Workshop of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics(HLTNAACL), Edmonton, Canada, 2003, pp: 43-48.

[11] Christiane Fellbaum. “WordNet-An Electronic Lexical Database”[M]. The MIT Press. 1998.

[12] Peng XP. “Research on the Technique of Topic Detection and Tracking of Microblog” [D]. Huazhong University of Science and Technology, 2013. (in Chinese with English abstract)

[13] Wu ZB. “Research of Detection and Tracking on the Chinese Micro-Blog Topic and System Design”[D]. Guangdong Polytechnic Normal University. 2013. (in Chinese with English abstract)

[14] Salton G, Buckley C. “Term-weighting Approaches in Automatic Text Retrieval”[C]. In: Int. Proc. of Mgt, 1988: 513-523.

[15] Li SP, Zhao J, Song ZC, etc. “Study on Topic Tracking System Based on SVM”[C]. In Proc. of 2011 Fourth International Symposium on Knowledge Acquisition and Modeling(KAM), IEEE Computer Society, 2011:83-87.[DOI:10.1109/KAM.2011.8]

[16] Zhou XG, Gao F, Sun Y. “Sub-topic detection and tracking based on dependency connection weights for vector space model”[J]. Journal on Communications, 2013, 34(8): 1-9. (in Chinese with English abstract) [DOI:10.3969/j.issn.1000-436x.2013.08.001]

[17] Li TT, Ji DH. “Sentiment analysis of microblog based on SVM and CRF using various combinations of features”[J]. Application Research of Computers, 2015,32(4):978-981 (in Chinese with English abstract). [DOI: 10.3969/j.issn.1001-3695.2015.04.004]

[18] Li YG, Zhou XG, Sun Y, etc. “Research and Implementation of Chinese Microblog Sentiment Classification”[J].Journal of Software. 2017, 28(12): 3183-3205. [DOI:10.13328/ j.cnki.jos.005283]