Research on New Event Detection Methods for Mongolian News

Shijie Wang, Feilong Bao✉, Guanglai Gao
College of Computer Science, Inner Mongolia University
Inner Mongolian Key Laboratory of Mongolian Information Processing Technology
National & local Joint Engineering Research Center of Mongolian Intelligent Information Processing Technology
Hohhot, China, 010021
997193900@qq.com; {csfeilong, csggl}@imu.edu.cn

***Abstract*—New event detection (NED) aims at detecting the first news from one or multiple streams of news stories. This paper is aimed at the field of journalism and studies the related methods of Mongolian new event detection. The paper proposes a method that combines the similarity of news content with the similarity of news elements to detect the new event. For the news content representation, according to the characteristics of the news and the different vocabulary expressions in different news categories, improve the traditional TF-IDF method. In addition, extract the main elements of the news, including time, place, subject, object, denoter, and calculate the similarity of news elements between the two news documents. Finally, the similarity between the news content and the news elements is combined to calculate the final similarity for new event detection. The experimental results show that the improved method is obvious, and the performance is significantly improved compared with the traditional new event detection system.**

Keywords-NED; Mongolian; News Elements; Similarity combination

I. Introduction

New event detection (NED) is defined as detecting the earliest report of a new topic in a news stream. A topic consists of a seminal event and subsequent directly related events or activities. An event is something that happens at a certain time and at a certain place. An example of a topic might be an explosion of an airplane. The first story on this topic is the story that first carries the report on the explosion of an airplane's occurrence. The other stories that make up the topic are those discussing the death toll, the rescue efforts, and the commercial impact and so on. A good NED system would be one that correctly identifies the article that reports the explosion of an airplane's occurrence as the first story. NED system has important research value. Identifying the first report of the topic through the NED system helps the tracking and detection of topic (TDT) system to mine the seminal event of the topic and establish an initial topic model. In addition, NED has important application value. With the increasing expansion of Internet resources, new events are often submerged in a large amount of daily information, which greatly limits people's timely grasp of important news. In particular, the business tracking of financial and stock markets, as well as areas such as national security and political analysis, there is a need for a mechanism to accurately and effectively capture important news events.

After the development in recent years, the new event detection technology of resource-rich languages such as English and Chinese is becoming more and more mature. However, the research on the detection of new Mongolian events is still in its infancy. There are few resources available in Mongolian, and there is a lack of publicly labeled corpus. In addition, Mongolian is an agglutinative language. The characteristics of word formation are different from those of Chinese and English. There are also problems such as inconsistent coding. This paper is aimed at the field of journalism and studies related methods for detecting new events in Mongolian. In recent years, with the increasing network resources related to Mongolian, the contents of news web pages on the Internet have expanded rapidly. It is difficult for users to obtain information of interest from complex topics quickly and accurately. The demand for detection and tracking of Mongolian is constantly increasing. Therefore, it is of great necessity to conduct research on the new event detection task in Mongolian.

The main method of NED is to calculate the similarity between each news story on hand and all the previous received stories. If all the similarities between them do not exceed a threshold, then the story triggers a new event. In addition, there are also systems that organize the previous news into news clusters, each news cluster corresponds to a topic, and compares the new news with the previous news clusters. The experiments in the literature [3,5] prove that the former method can get better results. The core problem of the former methods is well-modeled representation of news context. General method separately used TF-IDF model [1], the context vector [2], some key words which were extracted by named entity recognition (NER) system [10, 11, 12, 13] to represent the context. Papka [7] excludes NEs that frequently appear in the corpus from the text description, and gives the location class NE four times the weight of other features. Giridhar [13] describes the report as three vector spaces, which are vectors containing all features, vectors containing only NE features, and vectors excluding NE features, and compares the effects of three vector spaces on NED, but some reports work better without NE involvement. In terms of text representation, Yang [5] only selects the best relevant reports in the class to describe the topic based on the classification of the previous report categories. Brants [19] improves the incremental TF-IDF weight calculation method and uses the vector space model for text representation. Based on this, the Hellinger distance is used to match the text relevance. Xu et al. [20] applied a time window strategy to improve the single-pass clustering in NED task. Hong et al. [21] propose a new event detection model based on division comparison of subtopic. Cang et al. [22] proposed a temporal topic model (TTM) to divide the

978-1-7281-5014-7/19/$31.00 ©2019 IEEE

topics and news into some smaller events according to different time expressions describing the particular time of the event occurrence.

For the news, the five elements of the news are very important information. It summarizes the objective connection of the news events themselves and plays an important role in describing the characteristics of the news. While there is few efficient work about extraction of news factor including subject, denoter, object, time and location which can represent the news more precisely. Based on this idea, this paper extracts the news element time, place, subject, object, denoter, calculates the similarity of news elements and fuses the content similarity to detect new events.

The paper proposes a new event detection method combining content similarity and news element similarity. In the news content, two methods are used to optimize representation of news content. 1) According to a large number of corpora, the title and the first paragraph of a news can often express the main meaning of a news, giving special weight to the feature words appearing in the title or the first paragraph. 2) The literature [6] verified that different types of entities and part of speech have different sensitivity levels in different categories of news, using statistical results to optimize the weight of different categories of news for different feature words. In addition, news elements are also very important information for news. In addition to calculating the similarity of news content, the paper also extracts the main elements of news and calculates the similarity of news elements between the two news documents. To extract the factors efficiently, we applied the BiLSTM+CRF method separately on headlines and the first paragraph to extract subject, denoter and object, and on the content to extract time and location.

The remainder of the paper is organized as follows. Section 2 formally formulates the basic event detection model. Section 3 introduces our improved model. Section 4 introduces the experimental corpus, experimental process and experimental evaluation and results. Finally, Sect 5 summarizes the full text and look forward to the future work.

II. Basic Model

This section presents the basic NED model that is similar to what most current systems use. The paper builds on this model. A basic NED model consists of three parts, story representation, the similarity calculation, the NED process.

A. Story Representation

Firstly, we extract term to represent the text as a collection of items, and then construct the weight vector according to the weight of items. In the traditional vector space model, we consider a single word as the term extraction. We adopt incremental TF-IDF model to calculate the weight of words in the document, the incremental TF-IDF model updates the model every time period. At time period t, the model is updated as follows:

$$df_t(\mathrm{w}) = df_{t-1}(\mathrm{w}) + df_{D_t}(\mathrm{w}) \qquad (1)$$

Where D_t represents the news set in the t period, $df_{D_t}(\mathrm{w})$ represents the number of documents containing the feature w in D_t, and $df_t(\mathrm{w})$ is the number of documents containing the feature w up to time t , The paper sets D_t to 50, which means that each time period t contains 50 documents.

After the above processing, the weight of the feature w in a certain time period t is calculated as follows:

$$\mathrm{Weight(d, w)} = \frac{\log(tf(d,w)+1)*\log(\frac{Nt+1}{df_t(w)+0.5})}{\sum_{w\prime\in d}\log(tf(d,w')+1)*\log(\frac{(Nt+1)}{dft(w')+0.5})} \qquad (2)$$

Where N_t represents the total number of reports before the time period t, and tf (d, w) is the number of times the feature w appears in the news d.

Each news d at time t can be described as:

d→{weight(d,w_1), weight(d, w_2),…,weight(d, w_n)}

Where n is the number of characteristic words of news d, and weight (d, w) represents the weight of feature word w in news d at time t.

B. Similarity Calculation

The paper uses the Hellinger distance to calculate the similarity between news content. For the two news d and q, the similarity between them is expressed as:

$$\mathrm{Sim(d, q)} = \sum_{w\in d,q}\sqrt{weight(d,w)*weight(q,w)} \qquad (3)$$

C. New event detection

The new document q at the time t will be compared to all the previous documents d. The maximum similarity is compared with the specified threshold. If it is larger than the threshold, the described event is considered to be the old event, and conversely, the described event is considered to be a new event.

III. Improved model

The basic model uses the vector to represent the news and then calculates the content similarity to determine whether the news is new or old. Then for the reasonable representation of the news content, optimizing the weight of the feature words is a way to improve the NED effect. Literature [6] verified that different types of news have different sensitivity to different types of entities and part of speech. The literature [6] uses statistical results to optimize feature word weights. This paper uses this method to the Mongolian news corpus for verification and optimization of feature word weights. Also, the first paragraph and the title of a news can often indicate the main meaning of a news. We will give special weights to the words that appear in the title or the first paragraph, and use the above two methods to better represent the news content. In addition, analyzing the characteristics of news, the five elements of news, who, when, where, why, and what are important information for the news, and contribute to the judgment of whether the two news belong to the same topic. We extract the news elements, calculate the similarity, and combine the content similarity as the final similarity to detect the new events.

A. Improve TF-IDF model according to news characteristics

The title and the first paragraph of a news can usually express the main meaning of the news report. Set the parameter ω to adjust the weight of the words appearing in the title or the first paragraph.

$$weight_A\ (\mathrm{d, w}) = \omega * \frac{\log(tf(d,w)+1)*\log(\frac{Nt+1}{df_t(w)+0.5})}{\sum_{w'\in d}\log(tf(d,w')+1)*\log(\frac{(Nt+1)}{dft(w')+0.5})} \quad (4)$$

TABLE I. EXPERIMENTAL RESULTS WITH DIFFERENT VALUES

Experiment	Miss/%	FA/%	Norm(C_{Det})
ω = 1.2	51.25	4.28	0.6654
ω = 1.4	**50.19**	**4.15**	**0.6362**
ω = 1.6	50.89	4.29	0.6628
ω = 1.8	51.43	4.34	0.6633
ω = 2.0	51.45.	4.37	0.6635
ω = 2.2	51.56	4.48	0.6679
ω = 2.4	51.89	4.69	0.6703
ω = 2.6	52.33	4.72	0.6721
ω = 2.8	52.24	4.51	0.6744
ω = 3.0	54.89	4.56	0.6758
ω = 4.0	55.26	4.59	0.6779

Through a lot of experiments, it is concluded that when the feature word appears in the title or the first paragraph, ω=1.4 has a better effect.

B. News description based on special weighting of different types of entities

The basic NED model judges whether it is a new event based on comparing the similarity of news content, then optimizing the assignment of feature word weights becomes an effective way to improve the effect of NED. According to the statistical results, different types of news have different preferences for different types of named entities. It is proposed to use statistical results to optimize the weight assignment of feature words to improve the effect of NED.

This article manually names the entity categories. Named entities include people, place name, institution name, date, time, currency, and percentage. This article focuses on the sensitivity of different named entities in different categories of news. Extracting named entities using the NER system [15]. Use the ℵ^2 statistical method to count the correlation between feature words and topics. For a feature word w and a topic t, first, get a dependency table:

TABLE II. DEPENDENCY TABLE OF FEATURE WORD W AND TOPIC T

News number	Belong to topic T	Not belong to topic T
Include w	A	B
Not include w	C	D

The statistical method of ℵ^2 is as follows：

$$\aleph^2\ (\mathrm{w, t}) = \frac{(A+B+C+D)*(A*D-C*B)^2}{(A+C)*(B+D)*(A+B)*(C+D)} \quad (5)$$

We use the nine news topics in the collected Mongolian news corpus to average the statistical results of the same category of named entities and the same category of topics:

$$\aleph^2{}_{avg}(P_k,\ R_m) = \frac{1}{|R_m|}\{\sum_{T\in Rm} P(w,t) * \aleph^2(w,t)\}\ ,\ k = 1 \dots K, m = 1 \dots M \quad (6)$$

Where K is the number of named entities categories (7 in this paper), M is the number of news categories (9 in this paper), P_k represents the set of feature words of the kth entity, R_m represents the set of topics of the mth topic category, p(w, t) is the probability that the feature word w appears in the topic T.

Improve the feature word weights as follows：

$$weight_B\ (\mathrm{d, w}) = \frac{weight(d,w)*\alpha^{\frac{class(d)}{type(w)}}}{\sum_{w'\epsilon d} weight(d,w')*\alpha^{\frac{class(d)}{type(w')}}} \quad (7)$$

Where type(w) is the named entity type of the feature word w, class(d) is the category to which news d belongs, and α_k^c is the weighting parameter corresponding to news category C and NER category K, setting α_k^c=$\aleph^2$ (k,C).

In the process of new event detection, this paper uses Boostexter [18] to pre-classify news according to nine topic categories. Boostexter is a Boosting-based machine learning algorithm, which learns a series of simple rules for building classifiers from training data. The feature word weight generated by the initial TF-IDF model is used as the classification feature. 1000 manual annotation data is used as the training set, and all the data are classified. The classification result is used to calculate the feature weight adjustment in Equation 7.

C. News element similarity calculation

1) News element extraction

This paper cites the method of [14], using Bi-LSTM+CRF model to extract the time, place, subject, object, denoter in Mongolian news. We extract the subject, object, and denoter that appear in the headline and the first paragraph, because not all subject, objects and denoter appearing in the entire news are beneficial to us, but only in the headline or the first paragraph is more effective and more close to the characteristics of the news.

a) Host-object similarity calculation

After the subject, object and denoter are represented by vectors, the cosine formula is used to calculate the similarity. Where a_i, b_i represents the word vector in the document, w_{ik} is the weight of the feature word in the document.

$$\mathrm{Sim(F1,F2)} = \frac{\sum_i(ai*bi)}{\sqrt{\sum_i ai^2 \sum_i bi^2}} = \frac{\sum_{k=1}^{n}(w_{ik}*w_{jk})}{\sqrt{\sum_{k=1}^{n} w_{ik}{}^2}\sqrt{\sum_{k=1}^{n} w_{jk}{}^2}} \quad (8)$$

b) Time similarity calculation

For Mongolian news reports, the expression of time is standardized as “XX XX XX Day”, so the time similarity of Mongolian news corpus is relatively simple.

$$\mathrm{Sim(Ti, Tj)} = \begin{cases} \frac{|T_i \cap T_j|}{|T_i \cup j|} & T_i \cap T_j \neq 0 \\ 0 & \text{Other} \end{cases} \quad (9)$$

c) Location similarity calculation

Since there is no complete gazetteer for the Mongolian place name level mark, we use the method of cosine similarity to match the similarity of the place by formula 8, and record it as Sim(L_i， L_j). For example: L_i indicates the location of the Mongolian news report d $_i$, L_j indicates the location of the Mongolian news report d $_j$. L_i= （ᠦᠪᠦᠷ ᠮᠣᠩᠭᠣᠯ, ᠬᠥᠬᠡᠬᠣᠲᠠ, ᠰᠠᠶᠢᠬᠠᠨ ᠲᠣᠭᠣᠷᠢᠭ) (means: Inner Mongolia, Hohhot, Sai Han District), L_j= （ᠬᠥᠬᠡᠬᠣᠲᠠ , ᠰᠠᠶᠢᠬᠠᠨ ᠲᠣᠭᠣᠷᠢᠭ, ᠵᠣᠣ ᠤᠳᠠ ᠵᠠᠮ) (means: Hohhot, Sai Han District, Zhaowuda Road), Loc =(ᠦᠪᠦᠷ ᠮᠣᠩᠭᠣᠯ, ᠬᠥᠬᠡᠬᠣᠲᠠ , ᠰᠠᠶᠢᠬᠠᠨ ᠲᠣᠭᠣᠷᠢᠭ, ᠵᠣᠣ ᠤᠳᠠ ᠵᠠᠮ) (means: Inner Mongolia, Hohhot, Sai Han District, Zhaowuda Road), use 0, 1 to determine if it is in Loc, L_i = （1110） , L_j=(0111), Calculate the similarity according to the cosine formula to 0.667.

The calculation process that makes the place name similarity error smaller is as follows: First, we need to establish a Mongolian location database, which is established according to the location level; secondly, when the location of the corpus is extracted, if the two names have the same word, use the formula (10) to calculate Similarity. If there is no identical word in the two place names, check the place name database to confirm whether the two places are the same place. If it refers to the same place, use the formula (10) to calculate the similarity of the place names. If it is not the same place, the two similarities are 0.

$$\mathrm{Sim}(L_i, L_j) = \begin{cases} \frac{L_i \cap L_j}{L_i \cup L_i}, & \text{if } L_i \text{ and } L_j \text{ have the same place name and alias} \\ 0, & \text{other} \end{cases} \quad (10)$$

2) Combination of similarity

After calculating the content similarity, time similarity, location similarity, subject, object and denoter similarity between news stories, we put them together to form the final similarity of the two news reports. The final similarity calculation formula is as follows:

$$\mathrm{Sim}\left(d_i, d_j\right) = \alpha * \mathrm{Sim}(\mathrm{d}, \mathrm{q}) + \beta * \mathrm{Sim}\left(T_i, T_j\right) + \gamma * \mathrm{Sim}\left(L_i, L_j\right) + \delta * Sim(F1, F2) \quad (11)$$

Here, α, β, γ, δ are parameters. In the experiment, α = 0. 7, β = 0.1, γ = 0.1, δ = 0.1.

After calculating the event similarity, we need to determine whether the event is a new event. This paper judges whether it is a new event by comparing with the set threshold. If the final similarity is smaller than the threshold, it is a new event. This article sets the threshold to 0.2.

IV. Experiment

A. Data preparation

This article obtains corpus from the Mongolian news website, and uses Lucene to collect and sort out nine types of news corpus, namely earthquake, explosion, fire, tourism, sports, election, finance, crime, science discovery. These nine categories are used as experimental corpora. There are 1200 news reports, of which 800 are training data and 400 are test data. In the study of this paper. Mongolian news was selected as the experimental corpus. These news corpora were derived from Mongolian web-sites such as District Love Network and Hohhot. These corpora downloaded from the website cannot be used directly in the experiment and need to be processed. In the study of this paper, the preprocessing of Mongolian news corpus includes code conversion, text proofreading, removal of stop words, and removal of affixes.

For the experiment, the work of corpus annotation is as follows:

A news corpus mainly includes the title and content, and we mark the title and content. Expand the corpus by segment and expand the segment by sentence. Mark the subject, object, time, place, denoter, people, place name, institution name, date, time, currency, and percentage in each sentence. At the same time, we also mark the specific relationship between the sentence and the sentence, including Causal, Accompany, Follow relationship and so on.

B. Experimental evaluation

The Miss Rate (Miss Rate, Miss) and the False Rate (False Rate, FA) are used as the basis for the evaluation. The cost function C_{Det} is used for evaluation. C_{Det} comprehensively considers Miss and FA, and the calculation formula is as follows:

$$C_{Det} = C_{Miss} * P_{Miss} * P_{target} + C_{FA} * P_{FA} * P_{non-target} \quad (12)$$

C_{Miss} and C_{FA} respectively represent the cost function of missed detection and false detection. The values of both are generally set in advance. According to the empirical value, C_{Miss}=1 and C_{FA}=0.1. P_{target} and $P_{non-target}$ respectively indicate the probability of occurrence of a new event and the probability of occurrence of a non-new event, P_{target}=1-$P_{non-target}$. P_{miss} indicates the conditional probability of the missed detection rate, and P_{FA}indicates the conditional probability of the false positive rate.

Use a standardized cost function as the final evaluation criteria:

$$\mathrm{Norm}(C_{Det}) = \frac{C_{Det}}{\min(C_{Miss} * P_{target}, C_{FA} * P_{non-target})} \quad (13)$$

The standardized cost function Norm (C_{Det}) was used as an indicator for experimental evaluation. For a system that is fully judged correctly, Norm (C_{Det}) is 0, and all news is judged as new event news.

C. Experimental design

In order to test the effect of the improved model proposed in this paper, the following five systems were implemented and tested:

SYSTEM-1, this system is the baseline system, using the basic model introduced in the second section, that is, using the incremental TF-IDF model to generate feature word weights, using weight (d, w) as the weight of the t-time feature w in the news. The weights are then calculated for content similarity for event detection.

SYSTEM-2 uses the news description method proposed in Section III A, that is, assigning special weights to the words appearing in the title and the first paragraph, using weight A (d, w) as the weight of the feature w in the news at time t And then calculate the content similarity for new event detection.

SYSTEM-3 adopts the method of Section III B, which gives different weights of different entities to different types of news, and uses weight B (d, w) as the weight of feature w in the news at t time, and calculates the content similarity for new event detection.

SYSTEM-4 adopts the methods in Sections III A and III B, that is, giving special weights to the words appearing in the title and the first paragraph, giving different weights of different entities to different types of news, and using the comprehensive adjusted weights to represent the news content. Calculate content similarity for new event detection.

SYSTEM-5 comprehensively adopts the methods of Sections III A, III B, and III C, that is, assigns special weights to the words appearing in the title and the first paragraph, and gives different weights of different entities to different types of news, and uses the comprehensive adjusted feature weights to represent the news content and calculate the content similarity. Then extract the news elements, calculate the location similarity, time similarity, subject similarity, object similarity, denoter similarity, and finally merge with the content similarity for new event detection.

The above five systems were trained on 800 Mongolian news corpora that we collected and manually labeled, and tested in another 400 articles. In order to test the performance of the proposed method, this paper establishes four systems for comparison on the above data:

SYSTEM-6 when comparing two news documents, calculate three similarities, corresponding to named entities, unnamed entities, and all feature words, and use this similarity as a feature to use the support vector machine classifier to judge the news "new" or “old”.

SYSTEM-7 uses division comparison of subtopics. It utilizes the structure of topic and proposes the idea of subtopic.

SYSTEM-8 Depending on the news category, choose to use a named entity or a non-named entity to calculate the similarity and remove frequent words within the category.

SYSTEM-9 uses the method of re-evaluation of the word element, which gives special weights to different named entities and part of speech in different news categories based on statistical results.

D. Experimental results and analysis

TABLE III. COMPARISON OF TEST RESULTS

Experiment	Miss/%	FA/%	Norm(C_{Det})
SYSTEM1	50.23	4.25	0.6454
SYSTEM2	50.19	4.15	0.6362
SYSTEM3	48.27	4.01	0.5978
SYSTEM4	45.77	3.82	0.5834
SYSTEM5	**42.82**	**3.73**	**0.5415**
SYSTEM6	49.16	4.10	0.6324
SYSTEM7	44.86	3.79	0.5803
SYSTEM8	47.15	3.95	0.5928
SYSTEM9	45.33	3.80	0.5812

For the new event detection method applied in the Mongolian news field, the following conclusions can be obtained from Tables III:

SYSTEM2, the special weighting of the feature words appearing in the title and the first paragraph is slightly better than the general TF-IDF method. SYSTEM3, for the special weighting of entities that appear in different news categories, it is better than the general TF-IDF method. SYSTEM4 combines the above two methods to improve the TF-IDF model, the effect is better, and the missed detection rate and false detection rate are further reduced. SYSTEM5 combines the first two improvements, and combines the main elements of the news, calculates the final similarity to detect new events, this method can achieve the best results, the minimum standardization cost is reduced by 0.1039 compared with the baseline system SYSTEM1.Compared with System7, which has the best result of its kind, system5 is reduced by 0.0388.Explain that the system improvement effect is significant.

V. CONCLUSIONS

This paper proposes an improved method for the detection of new events in the Mongolian news field. This method combines the similarity of news content with the similarity of news elements, and compares the final similarity to detect new events. In the presentation of news content, according to the characteristics of the news, special weights are given to the words appearing in the title and the first paragraph, and different weights are given to the characteristic words of different entities in different news categories. Improve the TF-IDF model with the above two methods to better represent the news content. In addition, using the BiLSTM+CRF method to extract the main elements of news, including time, place, subject, object and denoter, calculate the similarity of news elements between two news documents, and combine the similarity of news content with the similarity of news elements. Calculate the similarity after fusion to detect new events. At the minimum standardization cost, the proposed method is reduced by 0.1039 compared to the baseline system, and is reduced by 0.0388 compared to similar systems. The experimental results show that the improved method proposed in this paper has a significant improvement on the results of new event detection in the Mongolian journalism field.

In this paper, the Mongolian experimental corpus is relatively small, which will affect the results of the experiment to a certain extent. In the future, it will continue to collect more corpus, do more comprehensive news classification, and continue to enhance the experimental results.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Nos. 61563040, 61773224); Natural Science Foundation of Inner Mongolia (Nos. 2018MS06006，2016ZD06).

REFERENCES

[1] Allan J, Papka R, Lavrenko V. On-line new event detection and tracking[C]. Proceed-ings of the 21st Annual International ACM SIGIR Conference on Research and De-velopment in Information Retrieval, New York: ACM Press, 1998: 37-45.

[2] Linmei Hu, Bin Zhang, Lei Hou, etc. Adaptive online event detection in news streams[J]. Knowledge-Based Systems, 2017: 105-112.

[3] Allan J, Lavrenko V, Malin D, Swan R.: Detections, bounds, and timelines: Umass and tdt-3. In: Proc, of the Topic Detections and Tracking Workship(TDT-3), Vienna, 2000, 167-174,http://cirr.cs.umass.edu/pubfiles/ir-201.pdf.

[4] Juha M, Helena A M, Marko S. Simple semantics in topic detection and track-ing.Information Retrieval, 2004, 7(3-4) :347-368.

[5] Yang Y, Pierce T, Carbonell J.:A study on retrospective and on-line event detection. In: Croft WB, Moffat A, Van Rijsberge CJ, Wilkinson R, Zobel J, eds. Proc, of the SIGIR'98. Melbourne, 1998.28-36

[6] Kuo Z . A New Event Detection Model Based on Term Reweighting[J]. Journal of Soft-ware, 2008, 19(4).

[7] Papka R, Allan J. On-line new event detection using single pass clustering TITL E2: Tech-nical Report UM-CS-1998-021,1998.

[8] Lam W, Meng H, Wong K et al. Using contextual analysis for news event detection. Interna-tional Journal on Intelligent Systems,2001,16(4):525-546.

[9] Juha M, Helena A M,Marko S. Applying Semantic classes in event detection and track-ing//Proceedings of the International Conference on Natural Language Processing(ICON 2002). Turko,Finland,2002:175-183.

[10] Yang, J. Zhang, J. Carbonell, and C. Jin. Topic conditioned Novelty Detection. In Proceed-ings of the 8th ACM SIGKDD International Conference, ACM Press. 2002.

[11] M. Juha, A.M. Helena, and S. Marko. Applying Semantic Classes in Event Detection and Tracking. In Proceedings of International Conference on Natural Language Processing.

[12] M. Juha, A.M. Helena, and S. Marko. Simple Semantics in Topic Detection and Tracking. Information Retrieval,

[13] K. Giridhar and J. Allan. Text Classification and Named Entities for New Event Detection. In Proceedings of the 27th Annual International ACM SIGIR Conference, New York,,USA:ACM Press,2004:297-304.

[14] Gao Yao-Wen: New event detection for Mongolian news corpus [D]. Inner Mongolia Uni-versity, 2018.

[15] Wang Wei-Hua. Mongolian named entity recognition research [D]

[16] Thorsten B, Francine C, Ayman F.A system for new event detection//Proceedings of the 26th Annual International ACM SIGIR Conference. Toronto ,Canda:ACM Press,2003:330-337.

[17] Croft W B, Townsend S C,Lavrenko V. Relevance feedback and personalization :A lan-guage modeling perspective//Proceedings of

the DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries. Dublin, Ireland,2001:49-54.

[18] Schapire RE, Singer Y. Boostexter: A boosting-based system for text categorization. Ma-chine Learning,2000,39(2/3):35-168.

[19] Brants T, Chen F, Farahat A. A system for new event detection [C]. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. [S. l.]: ACM,2003: 330-337.

[20] Xu RF, Peng WH, Xu J, et al. On-line new event detection using time window strategy [C]. In: 2011 International Conference on Machine Learning and Cybernetics (ICMLC). IEEE, 2011: 1932-1937

[21] Hong Y, Zhang Y, Fan JL, et al. New event detection based on division comparison of subtopic [J]. Chinese Journal of Computers, 2008, 31(4): 1-9.

[22] Cang Y, Hong Y, Yao JM, et al. New event detection based on temporal topic model [J]. Intelligent Computer and Applications, 2011, (1): 74-78.