Similarity calculation based on Mongolian news corpus

Yaowen Gao [a], Feilong Bao [b]✉ and Guanglai Gao [c]

School of Computer Science and Technology, Inner Mongolia University, Hohhot 010021, China

[a]1803142674@qq.com, [b]csfeilong@imu.edu.cn, [c]csggl@imu.edu.cn

Abstract: Similarity calculation is an important part of new event detection and effective computation of text similarity can remove redundant information and improve the efficiency of users' query. The paper mainly studies the calculation of the similarity between the Mongolian news materials. Because of the non-standard Mongolian news corpus, the corpus needs to be preprocessed in order to deal with the later work, which can improve the efficiency. So first of all, it is necessary to preprocess the news corpus, including code conversion、text proofreading、stop-words removal and suffixes removal. Then the news messages are mapped to vectors with a vector space model and calculating similarity between the vectors by Cosine formula. Finally, we choose precision、recall、F-measure as evaluation standard to evaluate the experimental results. The results show that the experiment is better than the manual.

Keywords: Similarity, Mongolian, Vector Space Model.

1. Introduction

Text similarity is an important subject in Natural Language Processing research, which plays an important role in information retrieval, text classification, text mining and so on. Therefore, the calculation of content similarity to Mongolian events becomes very meaningful.

In recent years, many researchers have made some achievements in the research of the calculation of content similarity. Duo Jian Wu [2] used word2vec model to research Chinese text similarity. Allan J, Papka R [3] used the term frequency–inverse document frequency (TF-IDF) method to establish a vector model for the report and finally selected the single-pass method for clustering. Changnian Sun, Cheng Zheng, Qing sun Xia[5] proposed a method of similarity calculation of Chinese text based on potential Dirichlet distribution(Latent Dirichlet Allocation，LDA) and used Jensen－Shannon distance to compute text similarity.

The article mainly studies similarity calculation based on Mongolian news corpus. The structure of this paper is as follows: The first part is the introduction, which describes the background, meaning and development of content similarity. The second part is the methodology framework. The third part focuses on the preprocessing of Mongolian corpus. The fourth part mainly introduces computational similarity method, which consists of the text representation model and the method of similarity calculation. Finally, we summarize the whole article.

2. The methodology framework

In this paper, similarity means the degree of consistency between the news corpuses. The main flow chart of similarity calculation between Mongolian news materials is shown in Fig. 1 below. The news corpus needs preprocessing which includes code conversion、text proofreading、stop-words removal、affixes removal、data classification. After preprocessing, it is important to calculate similarity and Vector space model (VSM) is applied to the similarity calculation process.

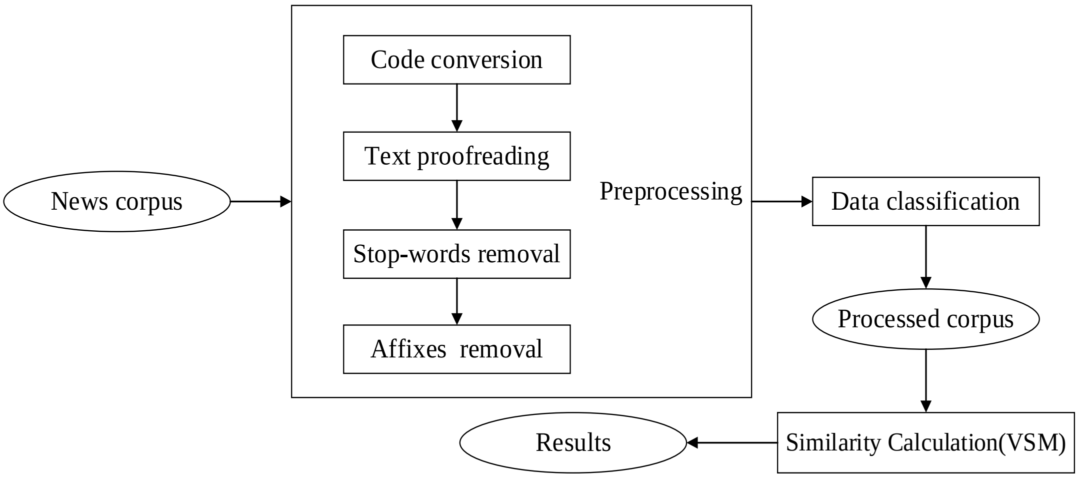

Fig. 1 Similarity calculation between Mongolian news materials

3. Preprocessing

The first of setting up corpus is to collect news corpus, followed by the need to preprocess corpus, which includes transcode corpus、cut affixes and other preprocessing operations. An example of text preprocessing is shown in Fig. 2 below. The meaning of this example is “I just want to talk to you both for a while”.

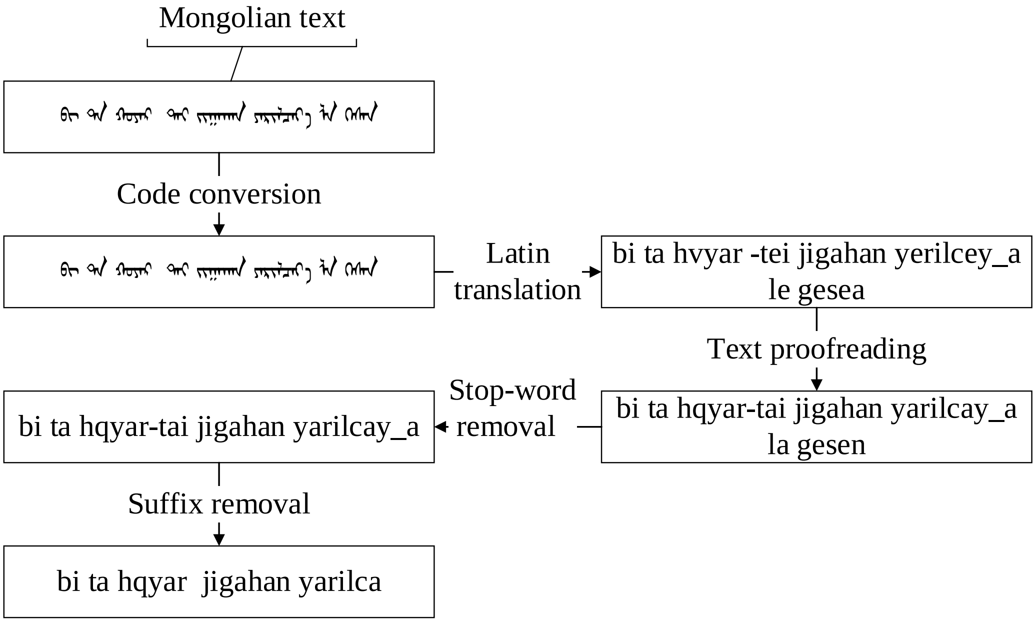
Mongolian text
Code conversion
Latin translation
bi ta hvyar -tei jigahan yerilcey_a le gesea
Text proofreading
bi ta hqyar-tai jigahan yarilcay_a la gesen
Stop-word removal
bi ta hqyar-tai jigahan yarilcay_a
Suffix removal
bi ta hqyar jigahan yarilca

Fig. 2 An example of text preprocessing

1）Code conversion

Due to the inconsistent coding of Mongolian corpus, it may happen that some Mongolian have the same writing style but different express meaning, which lead to the difference in internal coding. And the same character position will lead to different coding. The main content of this part is the conversion of the code, which transfer menksoft coding into international standard coding. The process of transcoding is to adjust the sequence of monk code according to the order of Mongolian encoding, so that they are corresponding one by one. Then we match each Mongolian single character in monk code with the corresponding encoding, in this loop until all characters of a word are found and convert it to the corresponding encoding coded string for display.

2）Text proofreading

Because of the grammatical features of Mongolian, there are a large number of words with different homonyms and different forms of homonyms. In addition, the formulation of the international standard for Mongolian coding and the promulgation and implementation of the "information technology of traditional Mongolian nominal characters, deformation characters and control characters using rules (GB25914-2010)" enforced by nation, so different control characters need to be used when inputting affixes, vowels, and special alphabet variants. This requires that the entry staff have a high grammatical knowledge of Mongolian and a serious attitude to accurately enter the Mongolian texts. However, at present, many people can't fully grasp these grammar knowledge and control rules. There are a lot of errors in Mongolian text input. After proofreading, it can be used as the original corpus for further research. This paper uses Mongolian automatic correction system to proofread the corpus. (The Mongolian automatic correction system is from http://mc.mglip.com:8080/.)

3）Stop-word removal

Stop words are words that have no practical meaning, such as auxiliary words, adverbs, prepositions, connectives and so on. For example, "ᠲᠤᠯᠠ、ᠪᠠ、ᠶᠤ、ᠪᠣᠯᠤᠨ" are Mongolian stop words. Removing the stop words is helpful to improve the search efficiency of documents. The specific operation is to refer to the Mongolian stop-word list, the word of the document and the stop word list are compared, if the stop word, it will be removed.

4）Suffixes removal

Mongolian belongs to agglutinative language, Mongolian word is composed of root suffix and a plurality of suffixes and all roots and suffixes can be combined into large-scale Mongolian words. The Mongolian suffixes can be divided into word-building suffixes, configuration suffixes, and ending suffixes. Root suffix or configuration suffix form stem, stem can add configuration suffix or ending suffix. The relationship between Mongolian roots, stem and suffix as shown in Table 1. The end suffix is a suffix that is at the end of the word only to express the simple grammatical meaning (mainly relation meaning). The basic meaning of the word itself does not change after the word is cut off at the end of the suffix. This paper segments the ending suffixes of Mongolian words and leaves Mongolian words stems, which can improve the search efficiency、reduce the confusion and reduce the storage space. The traditional Mongolian affixes need to be converted into Latin form firstly, and then the affixes are removed by referring to the suffix list.

Table 1 Relationship between Mongolian roots, stem and suffix

<table>
<tr><td colspan="3">stem</td><td></td></tr>
<tr><td>root</td><td>word-building suffix</td><td>configuration suffix</td><td>ending suffix</td></tr>
<tr><td></td><td colspan="3">suffix</td></tr>
</table>

4. Vector Space Model

4.1 Text representation

Vector space model is a simple and efficient text representation model. The vector space model is used to represent the text information into a vector form, so the text data can be converted into structured data that can be processed by the computer. To express the text as a vector, firstly, it is need to extract feature and text is represented as the set of feature items. Then we can use the TF-IDF method to calculate the weight of feature, which indicates the importance of the feature item to the text. Finally, the weight is arranged in descending order and the text is expressed by vector according to weight.

The text D is abstractly expressed as D= (t_i, w_i) =$(t_1, w_1, t_2, w_2 \ldots t_i, w_i)$, t_i denotes the feature term, that is the preprocessed word. W_i denotes the corresponding weight of t_i, that is the importance of t_i in the text.

4.2 Similarity calculation

The similarity calculation includes not only the similarity between the news reports and reports, but also the similarity between news reports and news class. The similarity is calculated by using the cosine formula. Specific analysis as follows:

1）The first step is to extract the key words that calculate the weight by TF-IDF. First of all, TF is calculated, that refers to the word frequency and the number of times a word appears in the article. For example, $TF_{i,j}$ represents the number of times a word numbered i appear in the article j. The second part is on computing the inverse document frequency (IDF). W_i is the weight of word i. The greater the W values of a word, the higher the importance of a word to the article. We calculate the weight by the following formula.

$$W_i = TF_{i,j} * IDF_i \tag{1}$$

The IDF_i formula is as follows:

$$IDF_i = \log \frac{\text{The total number of documents in the corpus}}{\text{the number of documents containing the word i +1}} \tag{2}$$

2）Using the cosine formula to calculate the similarity between the articles, the formula is as follows：

$$\text{Sim}\ (F_1, F_2) = \frac{\sum_i (a_i * b_i)}{\sqrt{\sum_i a_i^2 * \sum_i b_i^2}} = \frac{\sum_{k=1}^{n} (w_{ik} \times w_{jk})}{\sqrt{\sum_{k=1}^{n} w_{ik}^2 \sum_{k=1}^{n} w_{jk}^2}} \tag{3}$$

F_1, F_2 represents the document, a_i represents word vector in document F_1, b_i represents word vector in document F_2. w_{ik} is the weight of the word numbered k in the document numbered i.

5. Experiment and analysis

5.1 Experimental corpus

The corpus is derived from the people's network, the people's Government of the Inner Mongolia Autonomous Region and many other Mongolian websites. There are more than 350 reports are selected as a training set, including five types of events, namely earthquake, explosion, sports, tourism, Naadam, as shown in the following Table 2. About 600 reports were selected from the corpus, and 250 reports are randomly selected from the remaining corpus as a test corpus for a small scale experiment.

Table 2 Training set

News topic	The number of news reports
Earthquake ᠭᠠᠵᠠᠷ ᠬᠥᠳᠡᠯᠭᠡᠨ	30
Explosion ᠲᠡᠰᠡᠷᠡᠯᠲᠡ	25
Sports ᠲᠠᠮᠢᠷ	86
Tourism ᠵᠢᠭᠤᠯᠴᠢᠯᠠᠯ	175
Naadam ᠨᠠᠭᠠᠳᠤᠮ	54

5.2 Evaluation criteria

Three indicators of precision, recall ratio and F-Measure are mainly used as evaluation index. The value of precision and recall ratio is between 0 and 1, the closer the value to 1, the higher the precision or the recall ratio. F-Measure is a combination of evaluation indicators of these two indicators.

Through experiments, we get the result of recall and the comparison of the experiment and artificial in recall is shown in Fig. 3 below. The comparison of the experiment and artificial in F-Measure corresponding to different types of news is shown in Fig. 4 below.

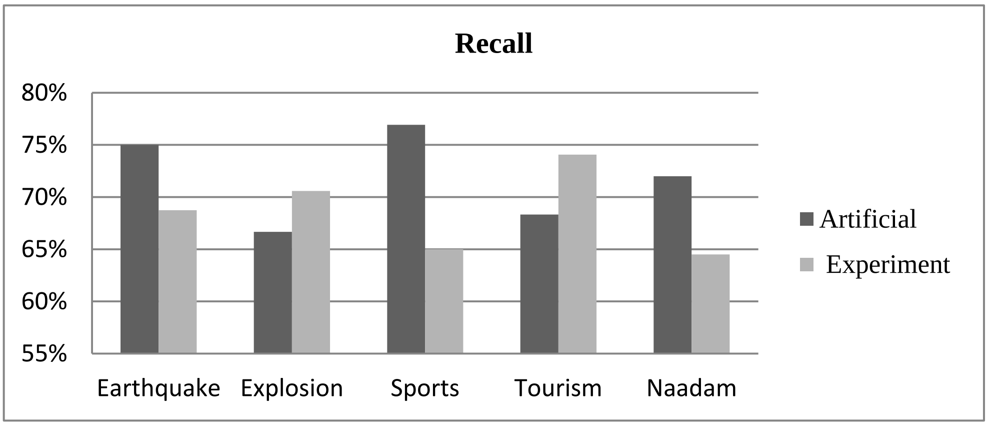

Fig. 3 Comparison of the two methods in recall

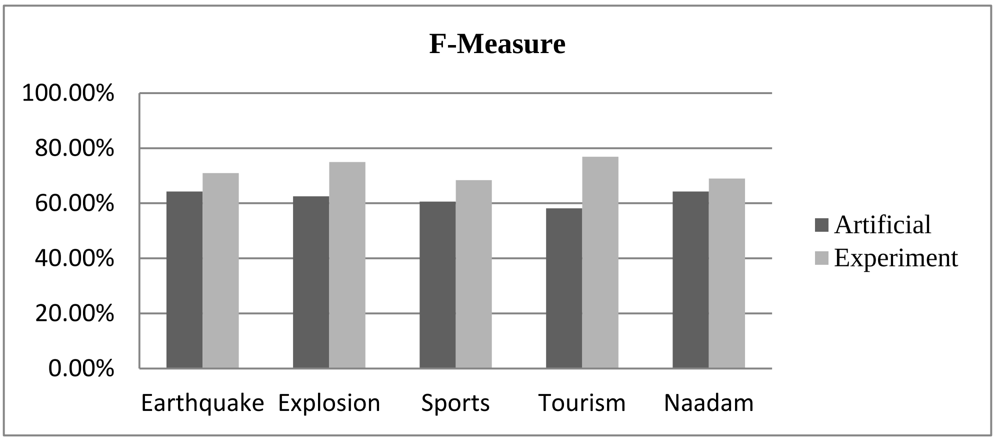

Fig. 4 Comparison of the two methods in F-Measure

6. Conclusions

From the evaluation results, we can see that the accuracy and F-Measure of this method are both higher than manual operation, so the experiment has better effect than manual processing.

The paper introduce the background of the similarity to the Mongolian news、existing research results. Then the target of the new event detection and the related techniques and methods used in the experiment are introduced. In the process of experiment, we need to preprocess corpus firstly, then we use vector space model to represent text, and finally we use cosine algorithm to calculate text similarity. Similarity calculation based on Mongolian news corpus still has a lot of research space. Moreover, similarity calculation based on Mongolian news corpus plays an important role in the new event detection based on Mongolian news corpus. In future research, we will apply the similarity calculation based on Mongolian news corpus to the new event detection for Mongolian news corpus.

7. Acknowledgments

This research was supports in part by the China national natural science foundation (No.61563040, No.61773224) and Inner Mongolian nature science foundation (No. 2016ZD06).

References

[1] Yue Wang. Topic Model Based Text Similarity Measure for Chinese Judgment Document [A]. ICYCSEE Steering Committee. Abstract of the Third International Conference of Pioneering Computer Scientists, Engineers and Educators, ICPCSEE 2017 PartII[C].ICYCSEE Steering Committee:,2017:3.

[2] Duo Jian Wu. Research and implementation of Chinese text similarity based on word2vec [D]. XiDian University, 2016.

[3] Allan J，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, New York: ACM Press,1998:37－45.

[4] Weiling Chen. Document Similarity Calculation Model of CSLN [A]. IEEE Beijing Section. Proceedings of 2014 IEEE 5th International Conference on Software Engineering and Service Science[C].IEEE Beijing Section: 2014:4.

[5] Changnian Sun, Cheng Zheng, Qing sun Xia. Similarity calculation of Chinese text based on LDA [J]. Computer technology and development, 2013, 23(01):217-220.

[6] Han Lu. Using concept semantic similarity for documents classification [A]. Hong Kong Education Society. Proceedings of 2013 International Conference on Information and Communication Technology for Education(ICTE 2013 VⅠ)[C].Hong Kong Education Society: 2013: 8.

[7] Xinpan Yuan, Jun Long. Near-duplicate document detection with improved similarity measurement [J].Journal of Central South University, 2012, 19(08):2231-2237.

[8] Jing-Jing Cui, Hong-yu NIE, Jia Du. New event detection based on sorted subtopic matching algorithm [J].Journal of Chongqing University (English Edition), 2013,(04):179-186.

[9] Duo Jian Wu. Research and implementation of Chinese text similarity based on word2vec [D]. XiDian University, 2016.

[10] Xuqin Fan. New event detection method based on word pair vector space model [J]. Computer engineering and Application, 2010, (12):123-125.