Word2vec and dictionary based approach for uyghur text filtering

Turdi Tohti, Yunxing Zhao and Winira Musajan
School of Information Science and Engineering, Xinjiang University, Faculty of Computer, Room 801, Science and Technology Building, No.666, Road Shengli, Urumqi 830046, P.R.China.
Email: turdy@xju.edu.cn

Abstract. With emerging of deep learning, the expression of words in computer has made major breakthroughs and the effect of text processing based on word vector has also been significantly improved. This paper maps all patterns into a more abstract vector space by Uyghur-Chinese dictionary and deep learning tool Word2vec, at first. Secondly, a similar pattern is found according the characteristics of the original pattern. Finally, texts are filtered using Wu-Manber algorithm. Experiments show that this method can get obvious filtering accuracy and recall of Uyghur text information improved.

1. Introduction

Information filtering refers to the technology of selecting user-wanted information and blocking unwanted or illegal information from large scale information flow. Information filtering has wide objectives such as text, audio, image and video formats in general, while usual (narrow sense) information filtering mainly focuses on text information.

With the rapid development of network technology, the more and more text data has been available for people and research issues. However, the openness, interactive and global characteristics of Internet made good and bad contents intermingled in spread. Many bad contented information such as violent, reactionary, pornographic or with obvious harmful emotional incitement brought troubles to user and the whole society, as well. Therefore, how to avoid bad information and ensure the healthy and sage networking environment has become an important research issue.

With increasing support to the development of western regions of China, the minority ethnic groups in this region also welcomed the fast economic, cultural and educational progress. Information processing technology for Uyghur and other minority languages in Xinjiang has stimulation under this excellent environment, too. In particular, Uyghur Web technology witnessed dramatic steps to make digitized text documents. Including the digitization of classic resources written on paper, quite a large volume of Uyghur text documents are being appeared in digital format [1]. Simultaneously, Uyghur information processing also faces the same or more problems that were encountered in Chinese and English information processing. Useful information is always hard to get from large, disordered information source, while the unhealthy information from webpages, text messages and other text carriers put serious threat on social order, national benefit in long run. Therefore, it is very important and urgent to find an efficient way of Uyghur text filtering.

Early text filtering techniques were mostly based on simple keyword matching and statistically filtering, which is simple in use and stable in effect. But these techniques fail to consider the overall content of document and diversity of word representations, only the documents which has the fixed key words are indexed, and often lead to mis-indexing or mis-filtering to some extent. The rise of deep
learning let the representation of words in computer has new breakthrough. Deep learning in text filtering is also playing more and more important roles and providing better effects. Chinese and English text filtering technology has been matured, while Uyghur text filtering research is still in its infancy. The unique language features of Uyghur had many existing technologies can’t be directly referenced, more studies still need to be further carried out.

This paper presents a Uyghur text filtering method combining the advantages of word2vec, Uyghur-Chinese dictionary and Wu-Manber filtering algorithm. Among them, word2vec is used for expanding the matching pattern, the transformation between the original and expanded patterns are realized by Uyghur-Chinese dictionary, and last filtering is conducted implementing Wu-Manber algorithm.

2. Related Works

2.1. Chinese and English Text Filtering
As early as 1958, Luhn proposed the idea of a "business intelligence machine". Then, in 1982, the concept of information filtering was first proposed by Denning who was able to identify emergent and regular mails in real-time e-mailing and thus expanded the scope of traditional information generation and information collection. In 1997, the information filtering was classified into adaptive filtering, batch filtering and shunt filtering in TREC1997, making the research on information filtering more specific [2].

It can be said that since 1950s, along with artificial intelligence, expert systems, machine learning and other fields of research, information filtering technology has acquired many experiences and breakthroughs. Information filtering for foreign languages such as English, already got very much advance. The concepts and techniques for each processing units are getting continuously improvements [3]. The study of Chinese text filtering started late in time. At present, most of them are working on feature extraction, similarity calculation, user template selection and related filtering algorithms and models in adaptive filtering system. Domestic research institutes including Tsinghua University, Fudan University, Chinese Academy of Sciences, Asian Research Institute of Microsoft and Harbin Institute of Technology first started their steps for text filtering and were gifted admirable push in this field [4].

2.2. Text Filtering Based on Pattern Matching
Multi-pattern matching received extensive attention in the field of string pattern matching. Common multi-pattern matching algorithms include Aho-Co-rasick algorithm, AC-BM algorithm, Wu-Manber algorithm, etc. Among them, Wu-Manber Algorithm is the most popular one [5] [6].

A Wu-Manber algorithm is the combination of the filtering with Boyer_Moore algorithms. It uses character group to expand the transferring effect of bad characters and filter the pattern string through comparing with hash table, so that the matching time is reduced. Thus, we no longer use single character for investigation of matching, instead, a group of B characters is considered for every matching step, and the moving distance of pattern string is determined according to the comparing result[7][8].

Wu-Manber algorithm first conducts preprocessing on pattern string set and establishes three tables-Shift tables, Hash table, Prefix table which are used to filter text documents later [9].

This paper takes the left-right custom of Uyghur into consideration during whole procedures of preprocessing, table building and filtering, and made some modifications on Wu-Manber [10].

2.3. Word2vec
Since Hinton etc. put forward the concept of Deep Learning in 2006, it has been widely concerned for machine learning and text filtering. Deep learning uses the basic architecture of neural networks, automatically extracts high level and more abstract feature of samples through combination of multi-layer perceptron. The extracted high level features usually get better results with common machine learning techniques in last step. Deep learning got welcoming in many research fields due to the universality of neural networks that various kinds of features are easily manipulated and integrated.
The convolutional Neural Network-CNN in image processing, Deep Neural Network and Restricted Boltzman Machine in speech processing are already accepted as popular and recommended tools [11].

Thanks to Google that developed the deep learning based open-source language modeling tool Word2vec in 2012, in which a word is represented in form of numerical vector, and has been extensively concerned in the field of natural language processing. The tool implements a continuous bag-of-word model, and computes the skip-gram structure of the word vector. It takes the text set as input and generates the word vector for each word by training. These word vectors can be applied to other natural language processing problems as a feature of word. For example, you can calculate the degree of similarity between two words according to the word vector. Using word vector also can avoid the phenomenon of "curse of dimensionality" [12].

Word2vec is the realization of neural network based probability language model, is the combination of probabilistic model with artificial neural network to improve the n-gram model. The purpose of the n-gram model is to obtain the conditional probability distribution of the nth word $w_n$ under the condition that the previous n-1 words $w_1, w_{n-1}$ were given. In other words, if there are D words in the dictionary, N-gram is hoped to learn the $A^{n-1}D$ probability distribution $P (w_n | w_1...w_{n-1})$, with context of $w_1...w_{n-1}$. Word2vec also has similar goal, but Word2vec assumes that $P (w_n | w_1...w_{n-1}) = f (w_1...w_{n-1}, w_n)$, where $f (w_1...w_{n-1}, w_n)$ is a function represented by a neural network.

![Figure 1. N-gram model.](image1)

![Figure 2. Continue bag of word mode.](image2)

It can be seen from Fig. 1 and Fig. 2 that for N-gram model, finding the probability of $w_n$ with known context $w_1 ..., w_{n-1}$ needs to visit the huge probability distribution table; While in continue bag-of-word model, the probability of this value can be simply calculated by the function $f (w_1...w_{n-1}, w_n)$. This function first maps the word as a word vector, then sums the word vectors, and finally calculates the value of $f (w_1...w_{n-1}, w_n)$ [13][14].

In addition, Word2vec uses fewer parameters to represent model than N-gram does, that makes Word2vec able to avoid over-fitting problem to some extent. The most important thing is that each word after training has got the corresponding word vector, and it can be considered that the vector is the projection of the word in a semantic space

3. **Word2vec and Dictionary Based Approach for Uyghur Text Filtering**

The text filtering system proposed in this paper consists the modules of Chinese corpus preprocessing, Uyghur Chinese translation, Word2vec based training and mode expansion, and the filtering at last. According to Word2vec, words in corpus should be separated by space, and we found the Stanford Chinese word segmenting tool is good for this task in preprocessing. In order to avoid the inference of unnecessary words in corpus, the stop words has been filtered out from the word table using the special filtering tool from Harbin Institute of technology.

As word2vec does not support Uyghur text pattern directly at present, we used Chinese as intermediate language for the selected Uyghur patterns according to user’s need. The Uyghur-Chinese dictionary developed by the Key Laboratory of Multi-lingual information processing of Xinjiang University is applied for translation between Uyghur and Chinese text patterns. The translated Chinese text patterns are processed with Word2vec, and the extended Chinese pattern is also translated into Uyghur for last filtering, simultaneously.
In the Word2vec based corpus training and pattern expanding module, the preprocessed Chinese text corpus is trained using Word2vec to obtain word vectors corresponding to the corpus. A pattern from user is extended using the word vectors to make more text patterns.

The extended patterns are processed by the modified Wu-Manber filter in last stage to select the appropriate patterns fit for user’s need from the large Uyghur text corpus. The structure of whole filter system is shown in Fig.3.

![Figure 3. Flow of Text Filtering Process.](image-url)

### 4. Experimental Results and Analysis

#### 4.1. Preparation of the Experiment

The training corpus for Word2vec contains 10 years text documents collected from the websites of TianShan, Yaxin, with total volume of 4.4G including the categories of politics, economy and technology. The Uyghur text documents to be filtered are also from the websites of TianShan, Yaxin and Renmin, and has volume of 1.2G. The experiments are carried out on Lenovo desktop computer with 4G RAM and Win7 Operating system.

In this paper, we use Word2vec to train corpus, and then use the corresponding four word vector model file to expand pattern, and finally use the same pattern to filter the same Uighur text information.

#### 4.2. Evaluation Methods

Recall and Precision are important terms always used in the field of natural language processing. Recall indicates the percentage of the detected relative items over the original ones, while precision refers the percentage of relative items over the detected ones in search. The former reflects how much useful items detected from total items should have been detected, and the latter indicates how much is useful among the detected ones [14]. The mathematical expression of Recall and Precision are as follows:

\[
\text{Precision}(p) = \frac{\text{number of correct filtered text}}{\text{number of filtered text}}
\]

\[
\text{Recall}(R) = \frac{\text{number of correct filtered text}}{\text{number of should be filtered text}}
\]
4.3. Experimental Results

The performance of the text filtering system was evaluated by calculating the Precision and Recall of the filtered Uyghur texts. In this experiment, we selected the patterns: کومپيۇتېر (computer), وۇرۇس (virus), ياغاچ ئات (trojan), تور بېخەتەرلىكى (network security), يوچۇق (holes). First, Word2vec is used to train the text corpus includes three categories such as with politics, economy and technology contents. Then, model expansion is done for the original model and filtering results are obtained using Wu-Manber algorithm. The comparisons of filter results with and without pattern expansion are given in Figure 4 and 5.

![Figure 4. Experiment results of un-expanded patterns.](image1)

![Figure 5. Experiment results of expanded patterns.](image2)

Through above experimental results we found that the proposed method has a large degree of improvement both in precision and recall than that of the traditional Wu-Manber text filtering. It should be noted that the precision and recall for technology category are higher than other two categories. Above results can let us to say that the proposed filtering system is superior to traditional Wu-Manber based text filtering? Also, larger correlation between the user-selected pattern and the training corpus yields better performance in filtering.

Finally, comparative experiments of different volume of corpus are carried out on the category of technology using the collected 3.5G Chinese text documents from Wangyi, Renmin, Sogou etc. The experiment environment, user-selected patterns, Uyghur text documents to be filtered is kept same. The 3.5G training corpus is randomly divided into 500M, 1G, 2G, the corresponding accuracy and recall rate are shown in Figure 6:

![Figure 6. Experimental results with different volume of training corpus.](image3)
The experimental results indicate that, larger volume of training corpus can produce better filtering efficiency. This can be analysed as bigger corpus more possibly generates the extended pattern that is more similar to the original ones.

5. Conclusions and Prospects

The traditional pattern matching based text filtering has obvious disadvantages, though it is simple in implementation. First of all, user-selected pattern cannot accurately reflect the user’s filtering requirements; Secondly, selected pattern is much fixed and easily misses the synonymous word during filtering, so the recall and precision of filtering are influenced.

The word vector based filtering method proposed in this paper can further improve filtering performance using larger volume of training corpus and the vocabulary of Uyghur-Chinese dictionary. Therefore, still more modification on filtering system can be implemented. 1) to train more accurate word vector with larger training corpus. 2) To eliminate even more unnecessary words using more effective word segmenting tool and stop-word filtering tool. 3) To increase the volume of Uyghur-Chinese dictionary, and establish a Uyghur-Chinese dictionary of better perfection and large vocabulary. 4) To divide the user selected patterns into more specific classes to expand the pattern using the word vectors generated from the corpus of same category and increase of volume and precision of extended patterns. 5) To increase the number of layers of neural network to raise the semantic abstraction of word vectors and improve the Word2vec based language model.

6. Acknowledgments

This work has been supported by the National Natural Science Foundation of China (61562083, 61262062).

7. References

[1] Turdi Tohti et al. Semantics-based Feature Extraction and Its Application in Uyghur Text Classification. Journal of Chinese Information Processing, 28(4):140-144, 2014.
[2] Yalqin Almas et al. Uyghur Text Filtering Based on Vector Space Model. Journal of Xinjiang University (Natural Science Edition), 32(2):221-226, 2015.
[3] Gao Yang et al. Pattern-based topics for document modelling in information filtering. IEEE Transactions on Knowledge and Data Engineering, 27(6):1629-1642, 2015.
[4] Yang Wenchuan et al. Study of an improved text filter algorithm based on trie tree. IEEE International Symposium on Computer, Consumer and Control, pages 594-597, 2016.
[5] Zhang Hongli et al. An efficient parallel algorithm for exact multi-pattern matching. Security and Communication Networks, 9(8): 1688-1697, 2015.
[6] V Bhardwaj et al. A Comparative Study of Wu Manber String Matching Algorithm and its Variations. International Journal of Computer Applications, 132(17): 34-38, 2015.
[7] L Zhang et al. Improvement on Wu-manber multi-pattern matching algorithm. International Conference on Computer Science & Network Technology, pages 608-611, 2014.
[8] YJ Wang et al. A Word-length Based Wu-manber Multi-pattern Matching Algorithm. Journal of Chinese Computer Systems, 34(7): 1650-1653, 2013.
[9] MO De-min et al. An Improvement of Wu-Manber Multiple Patterns Matching Algorithm. Journal of Chinese Information Processing, 23(1): 30-34, 2009.
[10] Maimaitiyiming Hasimu et al. Unique Character Based Statistical Language Identification for Uyghur, Kazak and Kyrgyz. Journal of Chinese Information Processing, 29(2): 111-117, 2015.
[11] R Collobert et al. Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research (JMLR), 12(1): 2493-2537, 2011.
[12] EH Huang et al. improving word representations via global context and multiple word prototypes. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, pages 873-882, 2012.
[13] T Mikolov et al. Linguistic regularities in continuous space word representations. Proceedings of NAACL-HLT, pages746-751, 2013.
[14] WANG Yi-pei et al. An Improved Wu-Manber Multi-pattern Matching Algorithm for Chinese Encoding. Journal of Chinese Computer Systems, 36(4): 778-781, 2015.
[15] WANG Zhen et al. A Parallel Multiple Pattern Matching Algorithm on Chinese/English Mixing. Computer Engineering, 40(4): 318-321, 2014.