Ranking chat logs for crime investigation

Wen Xiao*

Jiangxi Police College, Nanchang, Jiangxi, China

Keywords: Chat logs, Short text analysis, Query ranking, Non-negative matrix factorization, Crime investigation.

Abstract. Aiming at the needs of criminal investigators to quickly find chat logs related to crime case, a ranking method for chat texts has been proposed. First, query keywords were used to search chat texts, and the feedback texts were divided into relevant texts and irrelevant texts with manual annotation, then non-negative matrix factorization (NMF) was utilized to obtain the implicit semantic relationship of chat texts, finally chat logs could be ranked based on the scores calculated by similarity. Experiments show that the method proposed in this paper can quickly retrieve and get the messages of interest in a lot of chat logs, which can facilitate crime investigation.

1 Introduction

With the popularity of mobile Internet, instant messaging software has become an important tool for people's daily communication, such as whatsapp, wechat, QQ, etc. The chat logs generated by those software contain a lot of information. In the process of crime investigation, the analysis of chat logs can be used to find investigative clues or direct evidence among many crime cases, which also turns the analysis of chat logs into an important part of the current case investigation process.

At present, there are some forensics software or data analysis tools on the market that can extract application data from various mobile phones, including application data of instant messaging software. When dealing with the collected data, the data analysis tools mainly focus on the analysis of call logs, organizer data, SMS, multimedia files, or deleted data etc [1]. There is still a lack of semantic analysis of chat message, and it is a difficult problem to quickly find relevant information about the crime case in chat logs [2]. Researchers have devised different methods to study how to extract chat information, such as feature selection [3], topic models [4], text mining [5], etc.

Chat message is a typical type of short text, which is the current research hot spot of text analysis [6,7]. People have conducted a lot of researches on short text. Among those they can be divided into two main types of methods: short text classification and short text clustering. The classification methods need to label the data so as to build a training data set, and then use various machine learning techniques to generate a classification model. However, the process of data labeling requires a large amount of manpower and money, moreover each crime case has its particularity, and therefore the training data set is generally difficult to

* Corresponding author: shauven@126.com

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

transfer, so this type of methods has certain limitations. The clustering methods are known as unsupervised techniques, which don't need to label data, but just classify the data into some categories according to the inherent characteristics of data, such as based on density of data, hierarchy of data, or partition of data. Yet the defects of clustering methods are obvious: the lower accuracy and the number of classes hardly to determine. More importantly, the clustering methods cannot conveniently contain the user's query requirements or intentions. In addition, various methods of information retrieval are also used in text analysis. The mainstream work can be divided into two directions: query expansion [8] and query re-ranking [9]. Query expansion is to expand the user's original query by combining the relevant documents with the user's feedback information, so that the new query can better describe the user's intent, and then use the new query to retrieve. The method of query re-ranking is based on the user's feedback information, re-ranking the original search results according to some principles to ensure that the documents required by the user are ranked at the forefront. According to the reality needs of case investigation and evidence collection, this paper proposes a ranking algorithm of chat logs based on non-negative matrix factorization and similarity calculation for crime investigation.

2 Main technology of short text analysis

The general process of text analysis is to preprocess the text first, which includes removing some special symbols, retaining meaningful text content, and then uses the word segmentation tool for Chinese text to perform word segmentation, meanwhile removes stop words in the document set. Now each word in the document set can be counted and each document can be represented as a vector. The classic method of vector representation is the TF-IDF representation, namely Term Frequency-Inverse Document Frequency, which is widely used in information retrieval and text analysis, specifically defined as:

$$tfidf(t,d) = tf(t,d)) \times \log\left(\frac{|D|}{df(t)}\right) \tag{1}$$

Where tf(t,d) means the occurrence number of the vocabulary t in the text d, df(t) means the number of texts that the term t appeared in the text set, and |D| is the number of texts in the whole text set. The term weight calculated by TF-IDF can reflect the term's ability to distinguish texts, that is, the larger the value, the higher the term's ability to distinguish texts, the more important the term is.

Different from general texts, short texts are flexible, diverse and highly redundant. Especially chat messages have the characteristics of loose structure, shortness, and contextuality, resulting in some issues such as the difficulty of extracting features and the sparsity of features in chat texts. It is inappropriate that utilizing the traditional TF-IDF vector to represent chat texts. In recent years, Non-negative Matrix Factorization (NMF) has been widely used in short text analysis due to its advantages in mining the implicit semantics of information [10]. NMF is a low-rank approximate decomposition model, which decomposes a matrix into a product of left and right non-negative matrices. The left matrix can be considered as a set of basis vectors, and the columns of the right matrix are the projections (or weight coefficients) on the set of base vectors corresponding to the columns of the original matrix. This combination form of base vectors has a very intuitive meaning, so that NMF not only has the ability of dimensional reduction, but also has good interpretability.

Given any non-negative matrix $V=(v_{ij})_{m\times n}=[v_1,v_2,...,v_n]$, the NMF algorithm looks for two non-negative matrices $W=(w_{ij})_{m\times r}$ and $H=(h_{ij})_{r\times n}$ such that $V\approx W\cdot H$, which r satisfies $(n+m)\cdot r<nm$. Non-negative matrix factorization is essentially an NP problem, which can be

transformed into an optimization problem. The iterative methods are used to alternately solve W and H [11]:

$$W_{ia} \leftarrow W_{ia} \sum_{\mu} \frac{V_{i\mu}}{(WH)_{i\mu}} H_{a\mu} \tag{2}$$

$$W_{ia} \leftarrow \frac{W_{ia}}{\sum_{j} W_{ja}} \tag{3}$$

$$H_{a\mu} \leftarrow H_{a\mu} \sum_{i} W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}} \tag{4}$$

In text analysis, there are multiple ways to measure the similarity of two vectors: Euclidean distance, cosine similarity, Jaccard similarity coefficient, etc. This paper uses cosine similarity to calculate the similarity between vectors, as shown in Eq. 5.

$$sim(d_i, d_j) = \frac{d_i^T \cdot d_j}{\|d_i\| \cdot \|d_j\|} \tag{5}$$

In general, most of the existing search tools are keyword-based, and the search terms are provided by an investigator. The quality of the retrieved results depends on the search terms provided [4]. This method has two problems: one is that the text retrieved by keywords query may also contain a part of the text that user do not want, and the other is that keywords query can only express less information, which may not cover user's query intent, resulting in some important information missing from the query results. Therefore, keywords query can only be used as a preliminary process of retrieval. If we can annotate some of the retrieved text to express more clear query intent, we can divide these texts into relevant texts and irrelevant texts. The higher the similarity between the text to be ranked and the text annotated as relevant, the more likely it is the related text that the user needs to find, and vice verse . Based on this assumption, the ranking score of query texts is calculated as follows [9]:

$$score(d_i) = \underset{d_j \in D_r}{Max}\{sim(d_i, d_j)\} - \underset{d_k \in D_{nr}}{Max}\{sim(d_i, d_k)\} \tag{6}$$

where d_i represents the text to be ranked, D_r represents the set of relevant texts, and D_{nr} represents the set of irrelevant texts.

3 Algorithm description

Fig. 1 describes the analysis process of chat logs, which is explained as follows:

(1) Pre-process the chat log texts, including word segmentation, removal of stop words as well as emojis, links, voices, etc.

(2) Use specific keyword related to crime case to make a preliminary query, and retrieve some texts of chat logs that may be involved in the case.

(3) Relevant feedback is used to manually label the retrieved text, and the actual records involved in the case are labeled as "relevant", and the remaining records are labeled as "irrelevant." If too many texts are retrieved, it is sufficient to label a small number of more

typical texts.

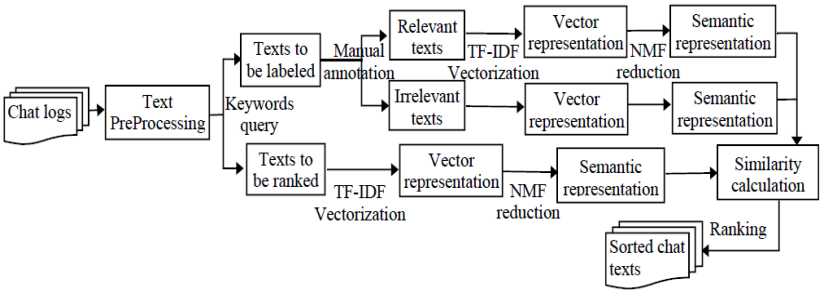

Fig. 1. Analysis process of chat logs based on query-ranking.

(4) Calculate the TF-IDF value for all texts, and represent the texts with TF-IDF vectors.

(5) Use NMF to reduce the dimensionality of all text vectors to obtain the low-dimensional semantic representation of the original vector.

(6) For all unlabeled text, use Eq. 6 to calculate the similarity score, and all unlabeled texts are ranked and displayed in descending order of score.

4 Experience results

In order to verify the effectiveness of the method proposed in this paper, the data used in the experiment comes from the WeChat chat logs of the suspects in a Internet MLM case. For the chat logs exported from the forensics software, the nearby chat messages are merged into one record and the time interval is set to 2 hours, the main purpose is to avoid the problem that the chat texts are too short and lacks context. Then some chat logs are manually labeled as positive and negative examples, a total of 1178 records are involved, of which 350 are related to the crime case. 87 relevant records and 45 irrelevant records are retrieved by keyword query and manually annotated in labeled chat records. The NMF-Similarity Method (NMF-S) described in this paper and the classic TF-IDF vector-similarity method (TFIDF-S) are used for comparison experiments. MAP (Mean of Average Precision), a common evaluation index for information retrieval, is used for evaluation, and P@N is used as an auxiliary evaluation index. P@N refers to the accuracy when N texts are retrieved for a certain user query, and MAP represents the arithmetic average of the query accuracy.

Table 1. Analysis results based on query-ranking.

	MAP	P@10	P@20	P@50	P@100	P@200
NMF-S	0.5887	1.0	0.95	0.96	0.88	0.605
TFIDF-S	0.4843	0.8	0.8	0.82	0.66	0.56

As can be seen from Table 1, the NMF-S method proposed in this paper has a significant improvement in query accuracy, with a MAP value of 0.5887, which indicates that NMF can effectively capture the semantic relationships implied in the texts and is more suitable for the analysis of short texts. The first 10 data of the experimental results are completely right, and the accuracy of the first 100 data can still reach 0.88, indicating that the method can quickly query and get most of the relevant information about crime case. It is noted that when more than 200 records are queried, P@N and MAP values have decreased significantly. The main reason is that when the chat messages are merged, the relevant information with crime case and some irrelevant information are merged into the same record, which leads the record to have a certain degree of similarity with relevant and irrelevant texts, therefore affects the final score.

5 Summary

An analysis method of chat logs for crime investigation is proposed in this paper, which is based on query and ranking. By means of keyword queries, a small number of chat records are labeled with relevant feedback. Relevant text sets and irrelevant text sets are obtained without additional burden for case investigators. Based on the similarity of texts, the rest of the chat logs are ranked. Experiment shows that the method can effectively seek, supplement and locate most relevant messages related to the case. Because many messages in the chat logs are too short, fixed time interval merging is used to overcome the problem of short text and lack of context. However, the processing method is likely to cause the information involved in the case to be mixed with other irrelevant information. Therefore, how to use the semantic relationship of context to merge chat messages is still a question that needs further research.

This research was financially supported by Collaborative Innovation Center for Economics crime investigation and prevention technology, Jiangxi Province (No. JXJZXTCX-019), and by Science and Technology Research Project, Jiangxi Provincial Department of Education (No.Gjj151196).

References

1. G. M. Jones, S. G. Winster. Forensics analysis on smart phones using mobile forensics tools [J]. International Journal of Computational Intelligence Research, 2017, 13(8): 1859-1869.
2. Y. Y. Wang, C. Q. Fan, Y. H. Su. Research on semantic analysis for chats [J]. Information Network Security, 2017 (9): 89-92.
3. S. Y. Li, W. He. Research on a method for feature selection of chat text [J]. Computer Science, 2007, 34 (5): 202-204.
4. A. R. M. Basher, B. C. Fung. Analyzing topics and authors in chat logs for crime investigation [J]. Knowledge and information systems, 2014, 39(2): 351-381.
5. K. Sameera, P. Vishwakarma. Cybercrime: To detect suspected user's chat using text mining [M]. Information and Communication Technology for Intelligent Systems. Springer, Singapore, 2019: 381-390.
6. H. Li, H. Huang, Cao X, et al. Falcon: A novel chinese short text classification method [J]. Journal of Computer and Communications, 2018, 6: 216-226.
7. J. Qiang, Y. Li, Y. Yuan, et al. STTM: A Tool for Short Text Topic Modeling[J]. arXiv preprint arXiv:1808.02215, 2018.
8. H. K. Azad, A. Deepak. Query expansion techniques for information retrieval: a survey [J]. Information Processing & Management, 2019, 56(5): 1698-1735.
9. B. Zhou, R. W. Cen, Y. Q. Liu, et al. A method for reordering search results based on document similarity [J]. Journal of Chinese Information Processing, 2010, 24 (3): 19-23.
10. C. B. He, Y. Tang, Q. Zhang, et al. Online clustering of short texts based on incremental robust non-negative matrix factorization [J]. Chinese Journal of Electronics, 2019, 47 (5): 1086-1093.
11. D. D. Lee, H. S. Seung. Learning the parts of objects by non-negative matrix factorization [J]. Nature, 1999, 401: 788-791.