Postprint: A Keyword Mining-Based Method for Screening Crime Leads in Hotline Text Data

Authors: Zhen Muhua, Chen Peng, Wang Kun, Fan Ziyang, king

Date: 2023-10-08T00:00:00+00:00

Abstract

[Purpose/Significance] To address the problem of insufficient information-based analysis capability in identifying and screening key information of crime clues from hotline text data in public security operations, this paper proposes a crime clue screening method for hotline text data based on keyword mining, which assists operational departments in improving the efficiency of intelligence analysis and assessment, thereby rendering crime clue screening work more information-based and scientific.

[Method/Process] Considering that the direct adoption of algorithmic methods such as text classification may lead to inadequate model training due to the excessively small proportion of valid information samples, this paper first extracts seed word sets from known crime clues based on text similarity, then utilizes Word2Vec to expand the seed vocabulary from two perspectives—synonymous words and alternative words—to construct a specialized lexicon, and finally achieves crime clue screening in hotline text data using a semantic-based scoring screening model.

[Results/Conclusion] By conducting crime clue screening experiments on 1,050 a priori hotline text data items from Jinan City and performing actual comparison and analysis of result metrics, a recall rate of 86% was obtained. It can be concluded that the semantic-based scoring screening method described in this paper achieves the expected performance in recognizing the specificity of crime information within Jinan City's hotline text data and realizes effective screening of crime clues.

Full Text

Research on Crime Clue Screening Methods for Hotline Text Data Based on Keyword Mining

Authors: Zhen Muhua[1], Chen Peng[1], Wang Kun[2], Fan Ziyang[1], Wang Zhe[1]
[1]School for Informatics and Cyber Security, People' s Public Security University of China, Beijing 100038
[2]Jinan Public Security Bureau, Jinan 250099

Abstract

[**Purpose/Significance**] Aiming to address the insufficient analytical capabilities in current public security operations for identifying and screening crime clues within hotline text data, this paper proposes a keyword mining-based method to help operational departments improve intelligence analysis efficiency and make crime clue screening more information-driven and scientific. [**Method/Process**] Considering that direct application of text classification algorithms may lead to inadequate model training due to the small proportion of valid information samples, this study first extracts seed word sets from known crime clues based on text similarity, then employs Word2Vec to expand the seed vocabulary from two perspectives—similar words and substitute words—to construct a specialized lexicon, and finally implements crime clue screening in hotline text data using a semantic-based scoring model. [**Result/Conclusion**] Experiments conducted on 1,050 prior hotline text data entries from Jinan City, with actual comparison and performance metric analysis, achieved a recall rate of 86%. This demonstrates that the proposed semantic scoring method meets expected performance for identifying specific crime information in Jinan' s hotline text data and enables effective crime clue screening.

Keywords: hotline text; specialized lexicon; text similarity; crime clue screening
Classification Codes: TP391; G250

---

1. Introduction

Hotline services represent an important initiative for public convenience, and hotline text data often serves as a crucial data source for public security agencies to investigate crime clues because it potentially contains intelligence information related to criminal activities. Currently, public security agencies predominantly employ a "tag system + manual screening" approach when processing hotline text data. Law enforcement officers first locate data categories that may contain key crime-related information through classification tags, then rapidly browse the detailed content fields and identify critical event information based on experiential knowledge, finally determining whether the data should be output as a crime clue. However, since detailed content fields typically appear as

long paragraphs where effective key information vocabulary units constitute a small proportion, extracting and mining critical information presents considerable difficulties, resulting in low effective analysis efficiency and inadequate data utilization in traditional manual screening modes.

The key to crime clue screening in hotline text data lies in recognizing and extracting key information representing criminal semantics within the text content. Existing research on key information extraction includes frequency-based keyword extraction methods (TF-IDF, LDA, etc.), which prove insufficient when key information vocabulary units account for a small proportion of the text. Additionally, in Chinese text analysis, frequency-based extraction suffers from semantic ambiguity issues. To address these limitations, researchers have proposed constructing key information lexicons using word vector technology (Word2Vec) combined with keyword extraction and text similarity calculation. For instance, Peng et al. used an SRC-LDA topic model based on semantic relationship constraints to extract theme words from product review texts; Liu et al. designed a sensitive lexicon using associated words and Jaccard coefficient expansion rules to retrieve and extract sensitive information from online public opinion texts, achieving over 10% improvement in reliability; Liu et al. constructed a photography domain sentiment dictionary using word vector models for sentiment information extraction and corpus classification; Tan et al. extracted disease feature information from cereal crop disease data; Xia et al. built a network rumor sensitive lexicon for short texts on social platforms like Weibo using Word2Vec-based semantic approximate matching; Tang et al. combined TF-IDF with word vector feature expansion for health question classification in medical Q&A data. These natural language processing techniques have achieved good results in various fields such as journalism.

However, public security agencies still rely on traditional manual screening for processing clue data due to issues like insufficient standardization in information expression and dispersed effective information. This paper takes hotline text data as an example, designs a method to extract crime clue key information based on its textual characteristics, and develops a semantic scoring model according to public security intelligence analysis logic to enhance the capability of obtaining crime clues from text data in an information-driven manner.

2. Methodology

2.1 Seed Lexicon Construction Word vector technology (Word2Vec) is a method for representing word meaning based on contextual distribution. It focuses on unlabeled data and uses neural network language models to learn semantic information from large texts, often employed to calculate similarity between words, sentences, or longer texts with good effect.

For seed lexicon construction, this study first collected crime information vocabulary from law enforcement departments as an experiential knowledge word set. Using the full dataset corpus as training data, Word2Vec models were built for

both the full dataset and the known attribute data (ordinary events/suspected crime clue events). The known attribute data word vectors served as the foundation for seed lexicon identification and extraction, while the experiential knowledge vocabulary word vectors acted as the matching reference set. Text similarity between them was calculated through vector mapping to extract information vocabulary meeting similarity requirements from the known attribute data, which was then collected to form the seed lexicon. The process is shown in [Figure 1: see original paper].

Extracted seed vocabulary falls into two categories: words representing suspected crime clue event semantics (Word_T) and words representing ordinary event semantics (Word_F). Here, "suspected crime clue events" refer to events that fall under public security agency jurisdiction according to relevant laws, including those with illegal behavior not reaching criminal standards, those requiring further confirmation, and those already filed needing supervision. Ordinary events refer to events not under public security agency jurisdiction, including invalid hotline events confirmed as malicious or repeated calls by relevant handling units.

To determine the reliability of extracted seed vocabulary in crime clue screening, we define a backtracking value as the ratio between the number of data entries (backtracking count) whose attribute is crime clue and contain a particular seed word, and the total frequency of that word in the full dataset. This represents the vocabulary' s reliability in crime clue screening:

$$P(word) = \frac{n(word)}{N(word)} \quad (1)$$

where $P(word)$ is the backtracking value, $n(word)$ is the backtracking count, and $N(word)$ is the word frequency in the full dataset. This backtracking value serves as the weight coefficient for the corresponding seed vocabulary in the crime clue screening model.

2.2 Lexicon Expansion Considering that the same semantics can be expressed through different vocabulary and sentence structures, we expand the lexicon from two aspects—similar words and substitute words—to achieve effective coverage. Combined with publicly available sensitive lexicons from the public opinion domain, this forms the extended word set. The reliability of extended vocabulary is determined by textual literal distance similarity between extended words and seed words, calculated using cosine similarity:

$$\text{Similarity} = \cos(\theta) = \frac{\vec{A} \cdot \vec{B}}{|\vec{A}| \times |\vec{B}|} \quad (2)$$

Similar Word Expansion. Word2Vec can reflect word context and semantic relationships. First, we obtain word vectors for the seed lexicon from the full

corpus Word2Vec model, then use the full dataset word vector model as the foundation for similar word identification. By calculating text similarity with the seed lexicon word vectors as the reference set, we extract key information words with similar semantics from the full corpus based on contextual relationships. The similarity score serves as the weight coefficient in the crime clue screening model. The process is shown in [Figure 2: see original paper].

Substitute Word Expansion. Considering that the same semantics can be expressed by different words, we use synonyms in Chinese expression as substitute words for the seed lexicon. Using the seed lexicon' s Word2Vec vectors from the full corpus combined with a Chinese synonym lookup tool, we find synonyms in an open-source Wikipedia Chinese corpus and calculate their text similarity to extract substitute vocabulary based on the public Chinese corpus. The similarity score serves as the weight coefficient in the crime clue screening model. The process is shown in [Figure 3: see original paper].

2.3 Crime Clue Scoring Model The scoring model is a public security data mining method emerging under the big data-driven intelligence-led policing model. This approach takes an event occurrence as the warning object, lists factors that may influence the event, and assigns corresponding weight scores according to their impact degree. Whenever a factor appears, a cumulative score is calculated until all factors are integrated. The integrated score represents a quantitative description of event occurrence probability:

$$\sum_{i=1}^{n} y_i \times p_i \quad (3)$$

where i represents influence factors, y represents score setting, and p represents the factor weight coefficient.

For hotline data research, the scoring value for a single data entry' s matching with a particular word set is influenced by three factors: the similarity value of matching individual vocabulary, the weight value of that vocabulary, and the count of vocabulary matching the same word in the word set. Additionally, publicly available sensitive words from the public opinion domain are only counted without duplicate scoring. The scoring rules for a single data entry against a word set are as follows:

$$S(dic) = a \times \sum_{i=1}^{n} \mathrm{sim}_i \times p_i + b \times \sum_{j=1}^{m} \mathrm{sim}_j \times p_j \quad (4)$$

$$\mathrm{SUM} = \sum S(dic) + \mathrm{Counts(internet)} \quad (5)$$

where $S(dic)$ represents the score for a particular word set type (seed, similar, substitute), S(Word_T) and S(Word_F) represent scores for word sets repre-

senting suspected crime semantics (T) or ordinary event semantics (F), a and b are weight coefficients for the word sets, SUM represents the total score, and Counts(internet) represents the count of non-duplicate vocabulary from publicly available sensitive word sets during matching.

2.4 Crime Clue Screening Algorithm Under the background where data has already been coarsely classified using a "tag system," hotline text data contains both event detail content information and invalid information such as punctuation and modal particles. Therefore, preprocessing is required before screening: using the Jieba Chinese word segmentation tool with a custom segmentation standard based on precise mode to avoid matching failures caused by different segmentation granularities; and removing punctuation and interference words using a custom stopword list.

This study employs a semantic-based scoring model for crime clue screening in hotline texts, where the total score of a data entry is composed of scores generated from matching with various word sets in the specialized lexicon. The screening process proceeds sequentially through three levels: similarity calculation between screening data vocabulary and lexicon vocabulary, scoring calculation between single data entry and a particular word set, and total scoring calculation between single data entry and the entire lexicon.

For vocabulary similarity calculation (match(seg, word)), which computes the similarity between a word (seg) in a screening data entry and a word (word) in a lexicon word set: (1) if the two words are identical, similarity is recorded as 1; otherwise proceed to (2); (2) if both words exist in the trained Word2Vec model, calculate their text similarity and proceed to (4), otherwise proceed to (3); (3) obtain seg' s word vector from the Wikipedia-based word vector model, calculate similarity, then proceed to (4); (4) if similarity meets or exceeds the threshold, record the similarity, otherwise end the calculation; (5) multiply the recorded text similarity by the weight value p of the matched word as the final similarity value.

For single data entry scoring against a word set (sim(data, dic)), the system iterates through elements in the input dataset for collision matching, combines the match(seg, word) module to sum similarity values, and simultaneously counts matching vocabulary occurrences in the word set. For total scoring and result output (sim(data, all)), after processing, the screening data entry obtains similarity values with all word sets. The scoring result is calculated according to the scoring rules designed in Section 2.1 and output. After completing a round of dataset screening, the screened data can be added to the database for dynamic updating.

3. Experimental Setup

The experiments primarily used the gensim.Word2Vec tool in Python 3.0 to construct word vector models. Experimental data came from 12345 Mayor Hotline

data provided by Jinan Public Security Bureau' s Food, Drug, and Environmental Crime Unit, covering January 2020 to March 2021, involving four domains: food and drug safety, medical supervision, environmental protection, and vaccine injection, totaling over 80,000 entries. Referencing actual public security business processes, the research data fields consisted of verified hotline event reply content from relevant administrative units, aiming to discover and supervise clues. Partial examples are shown in .

4. Experimental Results

4.1 Specialized Lexicon Construction Seed Lexicon. Following the seed lexicon construction method described in Section 2.1, we traversed experiential knowledge vocabulary in the training set, used the Word2Vec tool to calculate text similarity with known attribute data, and collected high-similarity vocabulary into the seed lexicon. Based on different attributes, the seed lexicon divides into two categories: seed_T representing suspected crime information semantics and seed_F representing ordinary event information semantics. The experiment yielded 94 seed words, with 55 in seed_T and 39 in seed_F, as partially shown in .

We further calculated backtracking values for generated seed words using formula (1) combined with stratified sampling. [Figure 4: see original paper] shows the relationship between word frequency and backtracking count for seed_T, while [Figure 5: see original paper] shows the backtracking value trend. For seed_T vocabulary, backtracking counts show obvious imbalance in proportion to word frequency, with backtracking values fluctuating irregularly. Overall, no clear correlation exists between backtracking values and word frequency, but the proportion of backtracking counts to word frequency reflects the characteristic that crime information occupies a small proportion in texts. Analysis reveals that since suspected crime semantics in seed_T mostly appear as short phrases, three types of words emerge after segmentation: conjunctions (e.g., "already"), neutral semantic words (e.g., "photograph," "investigate"), and terminology (e.g., "evidence collection," "suspect"). These three word types jointly determine crime information identification in texts, but conjunctions and neutral words alone cannot determine semantic nature and often appear with different terminology. When terminology appears alone, contextual judgment is required to determine if it represents crime semantics. Using word frequency as the identification standard for crime clue key information would significantly impact results.

[Figure 6: see original paper] shows the relationship between word frequency and backtracking count for seed_F, and [Figure 7: see original paper] shows the backtracking value trend. For seed_F, backtracking counts are proportional to word frequency, i.e., $n(word) \approx N(word)$, with backtracking values stabilizing mostly in the interval [0.8, 1). Unlike phrase-based information in seed_T, seed_F expresses ordinary event semantics, with phrases (e.g., "not included in assessment," "beyond jurisdiction") mostly composed of negative

conjunctions and terminology. When both appear simultaneously, the probability of classifying the data entry as an ordinary event is nearly 1, showing independent judgment capability. Meanwhile, most negative terminology also possesses independent judgment capability (e.g., “reject,”“malicious complaint”), making backtracking values of some negative vocabulary approach 1, indicating extremely high reliability for non-crime clue classification.

[Figure 8: see original paper] displays a directed network graph of segmented words in the seed lexicon, where node size represents word frequency and directed edges represent phrase structure relationships, with edge length determined by backtracking count. Larger nodes represent conjunctions or neutral semantic words, further demonstrating their lower reliability. Conversely, vocabulary clearly expressing suspected crime semantics appears as smaller nodes, with sentence structures mostly connecting to larger nodes, indicating higher reliability. Using backtracking values as influence factor weight coefficients in the scoring model can reduce result errors compared to using word consistency rules or frequency coefficient rules.

Extended Lexicon. For similar word expansion, we used Word2Vec to obtain mean vectors of the 94 seed words in the full corpus, then calculated similarity to extract 480 similar words: 251 in seed_T_similar and 229 in seed_F_similar, as partially shown in .

For substitute word expansion, after processing seed words with Word2Vec and combining with a Chinese synonym lookup tool, we extracted 506 substitute words: 271 in seed_T_synonym and 235 in seed_F_synonym, as partially shown in .

4.2 Crime Clue Screening Results The screening experiment used 1,050 data entries not involved in model training, including 1,000 ordinary event entries (Class F) and 50 suspected crime clue entries (Class T). Applying the proposed semantic-based screening method, we obtained scores for all entries, identifying 997 Class F and 53 Class T entries. Comparison with actual data revealed 43 correctly identified Class T entries. Performance metrics are shown in . Given the low proportion of Class T data, high recall was desired. The results show 86% recall and 81.13% precision, demonstrating that the proposed keyword mining-based scoring model achieves expected performance for crime clue screening in hotline text data.

5. Discussion and Conclusion

Effectively and scientifically extracting crime information from hotline data is crucial for law enforcement agencies processing text information and identifying crime clues. This paper proposes an automated screening method for crime clues in hotline text data based on keyword mining. First, a specialized lexicon is established through word vector models and text similarity calculation. Then, a scoring model for crime clue screening is designed based on the special-

ized lexicon. Empirical analysis using Jinan' s hotline text data demonstrates that this method can effectively identify specific crime information and screen crime clues, making screening work more information-driven and scientific. The method is also applicable to other public security text data processing tasks, such as public opinion monitoring.

Limitations include: (1) lexicon construction requires a certain number of experiential knowledge words and known target data samples; (2) future work could introduce paragraph vector models based on doc2vec for text classification, combined with qualitative weighted analysis using the specialized lexicon described herein.

References

[1] Wang Y. Application of big data in smart supervision of food and drugs in China[J]. China Food and Drug Administration, 2018(5): 44-47.
[2] Yuan M, Liu W, Hu J, et al. "Kunlun 2020" : Building a comprehensive food, drug, and environmental safety defense line[J]. People' s Public Security, 2020(16): 30-33.
[3] Xu J, Wang J, Ma W. Improved TF-IDF feature word extraction method using ontology correlation[J]. Information Science, 2011, 29(2): 279-283.
[4] Peng Y, Wan C, Jiang T, et al. Product feature and sentiment word extraction based on semantically constrained LDA[J]. Journal of Software, 2017, 28(3): 676-693.
[5] Liu G, Fang Y, Liu J. Sensitive lexicon design based on associated words and expansion rules[J]. Journal of Sichuan University (Natural Science Edition), 2009, 46(3).
[6] Liu Y, Lu X, Deng K, et al. Construction method of sentiment dictionary for photography domain reviews[J]. Computer Engineering and Design, 2019, 40(10): 3037.
[7] Tan M. Research on cereal crop disease identification and personalized push based on knowledge graph[D]. Changsha: Hunan Agricultural University, 2018.
[8] Xia S, Lin R, Liu K. Research on construction of network rumor sensitive lexicon: Taking Sina Weibo rumors as an example[J]. Knowledge Management Forum, 2019, 4(5): 267-275.
[9] Tang X, Gao H. Research on health question classification based on keyword vector feature expansion[J]. Data Analysis and Knowledge Discovery, 2020, 4(7).
[10] Jiang T, Wang S, Xu W. Chinese text classification based on naive Bayes[J]. Computer Knowledge and Technology, 2019, 15(23): 253-254, 263.
[11] Wu S. Research on basic issues in key personnel scoring warning model construction[J]. Journal of People' s Public Security University of China (Natural Science Edition), 2012, 18(2): 76-79.
[12] Tu M, Liu X, Liu S. Python Natural Language Processing: Core Technologies and Algorithms[M]. Beijing: China Machine Press, 2021: 120, 129.
[13] Yan H. Review of word vector development[J]. Modern Computer (Profes-

sional Edition), 2019(8): 50-52.
[14] Chen K J, Ma W Y. Unknown word extraction for Chinese documents[C]//Proceedings of international conference on DBLP. Taipei: Morgan Kaufmann Publishers, 2002: 169-175.
[15] Pedersen T, Kulkarni A. Identifying similar words and contexts in natural language with sense clusters[C]//Proceedings of the 20th national conference on artificial intelligence. Pittsburgh: AAAI Press, 2010: 1694-1695.
[16] Neviarouskaya A, Prendinger H, Ishizuka M. SentiFul: a lexicon for sentiment analysis[J]. IEEE transactions on affective computing, 2011, 2(1): 22-36.

Author Contributions

Zhen Muhua: Designed research methodology, conducted experiments, drafted and revised the manuscript.
Chen Peng: Proposed research ideas, revised the manuscript.
Wang Kun: Provided data, proposed research questions.
Fan Ziyang: Collected data, conducted experiments.
Wang Zhe: Collected data, conducted experiments.

Note: Figure translations are in progress. See original paper for figures.