Research and Application of Decision Tree-Based Sensitive Word Variant Recognition Algorithm (Postprint)

Authors: Dunhui Yu, Zhang Xiaoxiao, Fu Cong, Zhang Wanshan

Date: 2019-04-01T00:00:00+00:00

Abstract

To address the issue of inefficient identification of sensitive word variants in online environments, a decision tree-based sensitive word variant recognition algorithm is proposed. First, sensitive words and their variants are studied by analyzing characteristics such as the structure and pronunciation of Chinese characters. Second, a sensitive word decision tree is constructed based on a sensitive word lexicon. Finally, the sensitivity degree of text from new media platforms such as Weibo is calculated through a multi-factor improved model. Experimental results demonstrate that the algorithm achieves maximum recall and precision rates of 95% and 94%, respectively, in recognizing Chinese sensitive words and their variants. Compared with the improved algorithm based on deterministic finite automata, the recall and precision rates are increased by 19.8% and 21.1%, respectively; compared with the sensitive information decision tree filtering algorithm, the recall and precision rates are increased by 17.9% and 18.1%, respectively. Analysis shows that the algorithm is effective for the recognition and automatic filtering of sensitive word variants.

Full Text

Preamble

Vol. 37 No. 5
Application Research of Computers
ChinaXiv Partner Journal

Research and Application of a Sensitive Word Deformation Recognition Algorithm Based on Decision Trees

Yu Dunhui[1],[2], Zhang Xiaoxiao[1]†, Fu Cong[1], Zhang Wanshan[1],[2]
(1. College of Computer & Information Engineering, Hubei University, Wuhan

430062, China;
2. Education Informationization Engineering & Technology Center of Hubei Province, Wuhan 430062, China)

Abstract: To address the low efficiency in recognizing sensitive word deformations in online content, this paper proposes a sensitive word deformation recognition algorithm based on decision trees. First, the algorithm analyzes sensitive words and their deformations by examining Chinese character structure and pronunciation features. Second, it constructs a sensitive word decision tree based on a sensitive word lexicon. Finally, it calculates the sensitivity level of text from new media platforms such as Weibo using an improved multi-factor model. Experimental results demonstrate that the algorithm achieves maximum recall and precision rates of 95% and 94%, respectively, when identifying Chinese sensitive words and their deformations. Compared with an improved deterministic finite automaton-based algorithm, recall and precision improve by 19.8% and 21.1%, respectively. Compared with the sensitive information decision tree filtering algorithm, recall and precision improve by 17.9% and 18.1%, respectively. Analysis confirms the algorithm's effectiveness in recognizing and automatically filtering sensitive word deformations.

Keywords: sensitive word recognition; sensitive word deformations; decision tree; sensitivity computation; multi-factor model

CLC number: TP391.1
doi: 10.19734/j.issn.1001-3695.2018.11.0792

---

0 Introduction

With the rapid development of the Internet, online information has grown exponentially, and illegal speech (such as pornography, gambling, drugs, terrorism, and violent content) frequently permeates cyberspace [1,2]. This harmful information typically contains sensitive vocabulary, often appearing in large quantities as deformations, causing significant harm to the public, especially adolescents, and posing serious threats to national security, social stability, and a healthy online environment. Weibo, as a new broadcast-style social networking platform, widely disseminates, shares, and accesses brief information in real-time. However, due to its massive user base and limited regulatory capacity, malicious actors frequently distribute sensitive words through this channel. Therefore, identifying and filtering sensitive words and their deformations in new media platforms like Weibo has become an urgent research priority.

Numerous scholars have conducted research on sensitive word recognition and filtering. Reference [3] proposed analyzing text-sensitive words by constructing a CNN-like word network, which improved detection accuracy but required manual construction of word associations. Reference [4] introduced an improved deterministic finite automaton algorithm that builds a sensitive information decision tree using the first letters of sensitive word pinyin. While this approach

does not rely on a sensitive information corpus and improves detection efficiency, it cannot handle word deformations. Reference [5] proposed a sensitive information decision tree filtering algorithm that similarly improves search speed through decision tree construction and achieves sensitive text detection by assigning weights to sensitive words. However, this method relies on manually determined sensitivity levels, making it difficult to objectively represent text sensitivity, and lacks analysis and recognition of sensitive word deformations. Reference [6] presented a method for variant sensitive words that converts special characters into visually similar letters before detection, but its recognition efficiency for variant deformations remains low. Reference [7] employed machine learning methods using bigram and stemming features for text classification to detect deformations. These methods work well for English characters but do not account for Chinese sensitive word deformations.

In summary, current research on Chinese sensitive word deformation recognition and filtering suffers from insufficient analysis of deformations and low recognition and filtering efficiency. To address these issues, this paper proposes the **Recognition of Sensitive Words based on Decision Tree (RSWDT)** algorithm to tackle sensitive word deformation recognition and filtering. First, based on Chinese character pronunciation and structure, we analyze three deformation patterns: pinyin mode, abbreviation mode, and character splitting mode. Second, we expand the existing sensitive word lexicon by adding pinyin, location codes, and split location codes for words, then construct a sensitive word decision tree for accurate deformation recognition. Finally, combining an improved multi-factor model, we calculate text sensitivity for online content from Weibo, blogs, and comments to enable automatic filtering of sensitive text.

1.1 Analysis and Processing of Sensitive Word Deformations

This study examines three types of sensitive word deformations: pinyin mode, abbreviation mode, and splitting mode [8]. As online platforms increasingly scrutinize content, sensitive words in web text frequently appear in deformed versions, including these three patterns. Using the term “drug trafficking” (贩卖毒品) as an example, its deformation structures are illustrated in [Figure 1: see original paper].

1) Pinyin Mode
In Chinese, the same pinyin can correspond to multiple different characters. As shown in [8], this paper adopts a three-part phonetic code system that represents Chinese character pinyin as initials, finals, and tones, with each component encoded using English letters (a, b, etc.). This converts a Chinese character into a character sequence (phonetic code) for subsequent computation and comparison.

2) Abbreviation Mode
Statistics show that approximately 20% of sentences in Chinese news articles may contain abbreviations [9], including initialisms and word abbreviations. Ini-

tialisms include examples like “法轮功” abbreviated as “flg”. Word abbreviations generally fall into three forms: compression, truncation, and generalization [10], with compression and truncation being the most common combination. Compression divides the full term into several words and extracts the most representative characters from each, such as “贩卖毒品” abbreviated as “贩毒”. Truncation directly omits some words from the full term, retaining the remainder, such as “复旦大学” abbreviated as “复旦”. Both methods select partial characters or words from the full term to form abbreviations, generally preserving character order. Since all characters in an abbreviation are contained within the full term, finding a subset of the full term reveals its abbreviation.

3) Splitting Mode
Chinese characters can be categorized as single-component characters (e.g., 日, 月) or compound characters (e.g., 休, 取) based on their structural composition. Single-component characters consist of strokes, while compound characters consist of radicals. Spatial relationships between character components include intersection, separation, and connection [11]. Positional relationships include top-bottom, left-right, inside-outside, frame, and single-component structures. Location codes are four-digit decimal numbers, each corresponding to a unique Chinese character or symbol. Based on these characteristics, we manually split characters in the sensitive word list and encode them using location codes to create a character splitting table, as shown in .

To recognize split deformations of sensitive words, we first split both the sensitive word and the suspected deformation according to the character splitting table and convert them into corresponding location codes.

1.2 Overall Scheme

To achieve automatic sensitive information filtering, we implement the following steps:

a) Sensitive Word Deformation Recognition Based on Decision Trees
For the three deformation types (pinyin, abbreviation, and splitting), we propose a decision tree-based recognition algorithm that involves lexicon lookup, decision tree construction, and identification using the decision tree.

b) Automatic Sensitive Information Filtering Based on Multi-Factor Model
Based on recognized sensitive words and their deformations, we calculate text sensitivity by considering factors such as position, frequency, and category of sensitive words in the text, using an improved multi-factor model. Texts are then processed according to their sensitivity levels to achieve automatic filtering.

2.1 Sensitive Word Decision Tree Construction

Through analysis of sensitive words and their deformations, we recognize that identifying deformations requires phonetic, phonetic code, and location code analysis for each character. Therefore, before decision tree construction, we expand the existing sensitive word lexicon () to store pinyin, phonetic codes, and location codes for characters in known sensitive words, facilitating decision tree establishment and storage.

The decision tree construction algorithm classifies sensitive words by the first character's pinyin initial. It further clusters words with the same initial character, placing sensitive words with identical first characters under one branch to store each unique character only once, thereby improving retrieval speed and saving storage space. Each node stores the Chinese character along with its pinyin, phonetic code, and corresponding location code. Leaf nodes record the positions and categories of recognized sensitive words or deformations. The indices for position and category information in leaf nodes follow Huffman-like coding rules, using actual branch numbers when branch counts exceed two. The decision tree recognition algorithm can detect pinyin, abbreviation, and splitting deformations, such as "安 mian 药" or "亻乍".

Taking the sensitive word lexicon as input, the algorithm outputs a sensitive word decision tree, as shown in [Figure 2: see original paper].

The algorithm execution process is described in Algorithm 1.

Algorithm 1: ESDT Algorithm
Input: Sensitive word lexicon
Output: Sensitive word decision tree

a) Initialize and record child node indices.

b) Input sensitive word, obtain its Chinese character length and first letter.

c) Enter subtree query, compare with child nodes. If values match, continue; otherwise, check sibling nodes.

d) If sibling nodes exist, proceed to next step; otherwise, create new node.

e) Continue processing until all characters are stored, then create leaf node recording position and category.

f) Process next sensitive word until lexicon is exhausted.

g) Algorithm ends.

The constructed decision tree depth equals the length of the longest sensitive word in the lexicon (generally <10). Each node stores the character, its pinyin, phonetic code, and location code, while leaf nodes additionally record position and category information.

2.2 Sensitive Word Deformation Recognition

To accurately identify sensitive words and their deformations, the algorithm first obtains text containing suspected deformations, then inputs it into the decision tree. Starting from the first letter branch, it compares suspected characters with decision tree node information. When encountering Chinese-pinyin mixed suspected deformations (pinyin or abbreviation mode), direct matching occurs. When encountering special characters or radicals, location codes are obtained for matching in the splitting table. If matched, the algorithm proceeds to the next character; otherwise, it searches child and sibling nodes until reaching leaf nodes. Successful matches record positions and categories in leaf nodes. The process continues for all suspected words, finally outputting leaf node information containing positions and categories of sensitive words and deformations. The detailed execution is shown in Algorithm 2.

Algorithm 2: RSWDT Algorithm
Input: Sensitive word decision tree, text containing suspected deformations (where represents text characters and represents character count)
Output: Decision tree leaf node information, sensitive words and deformations

a) Initialize to record the first character sequence entering a branch.

b) Input text character , determine if it is Chinese, English, or a radical. For Chinese, extract first letter; for English, obtain directly; for radicals, obtain location code.

c) Match with child nodes; if no match, algorithm ends.

d) Record position in leaf node and output sensitive word or deformation, then return to step b.

e) If conditions are met, return to step d.

f) Check sibling nodes; if not empty, proceed to step c.

g) If sibling nodes are empty, return to step b.

h) Algorithm ends.

3 Multi-Factor Model for Sensitive Information Filtering

Multi-factor models are commonly applied in finance for quantitative investment decisions, combining multiple selected factors for final judgment. Here, we select position, category, and frequency of sensitive words in text as factors for calculating text sensitivity to enable automatic filtering. Using leaf node

information containing positions and categories of sensitive words and deformations output by the decision tree, we calculate text sensitivity through the following steps:

a) Output position information for each sensitive word from leaf nodes to form position information set.

b) Calculate position sensitivity for each sensitive word or deformation (frequency information is obtained through position accumulation).

c) Look up each word's category in the sensitive word type table. Each category has different weights, which are combined with position sensitivity to obtain each word's sensitivity.

d) Accumulate all word sensitivities to obtain overall text sensitivity, which assists in automatic text review.

3.1 Position Information Acquisition for Pinyin Mode

From the decision tree, we obtain the complete set of sensitive words and deformations along with their position information to form the sensitive word position information set , where represents the number of sensitive words and indicates each word's position in the text for calculating position sensitivity.

3.2 Position Sensitivity Calculation for Sensitive Words and Deformations

Due to information overload, people typically browse only the beginning and end of content to acquire maximum information quickly, consistent with the habit of placing summary descriptions at document heads and tails. Therefore, sensitive words appearing at the text head have greater impact on sensitivity than those at the tail, which in turn have greater impact than those in other positions. The position sensitivity of sensitive word is calculated as:

$$loc(s_i) = \begin{cases} \alpha & \text{if } 0 < l_i \leq a \\ \beta & \text{if } a < l_i \leq b \\ \lambda & \text{if } b < l_i \leq len(t) \end{cases}$$

where represents the position weight when sensitive word appears at the text head, middle, or tail; and are threshold values dividing the text into head, middle, and tail sections; and represents the position information of sensitive word .

3.3 Category Sensitivity Calculation for Sensitive Words and Deformations

Based on Xinhua News Agency's prohibited word regulations, sensitive words are categorized into five types: political and social life, laws and regulations, ethnic and religious affairs, Hong Kong/Macau/Taiwan and territorial sovereignty, and international relations. provides examples for each category.

Each category affects text sensitivity differently, requiring determination of relative weights among the five categories. This paper employs the Analytic Hierarchy Process (AHP) [12,13] to calculate these weights.

Assuming the category set where represents political and social life, represents laws and regulations, represents ethnic and religious affairs, represents Hong Kong/Macau/Taiwan and territorial sovereignty, and represents international relations, AHP determines the relative weight for each category.

3.4 Text Sensitivity Calculation

Using position, category, and frequency as factors selected by the multi-factor model, text sensitivity is calculated using Formula (2):

$$S(t) = \sum_{i=0}^{n-1} loc(s_i) \times typ(s_i)$$

where represents text sensitivity, represents position sensitivity of sensitive word , represents category sensitivity, and represents frequency.

Using normalization to map to the [0,1] interval, the normalized text sensitivity is:

$$S'(t) = \frac{S(t) - \min(S)}{\max(S) - \min(S)}$$

---

4 Experiments and Analysis

To validate the feasibility of the text sensitivity calculation method, we established an experimental environment, selected appropriate data, and conducted experiments under various conditions to collect and analyze results from multiple perspectives.

4.1 Experimental Environment

Experiments were conducted on a machine with a 2.4 GHz Intel(R) Core(TM) i7 processor and 8 GB RAM, running Windows 10. The programming tool was PyCharm, and the language was Python.

4.2 Dataset

To evaluate the Chinese sensitive word deformation recognition method, we downloaded 26,728 Weibo texts (covering technology, sports, finance, society,

entertainment, etc.) from CSDN (https://download.csdn.net) as the test dataset. After preprocessing, we manually identified and classified sensitive words and deformations, screening 3,835 texts containing deformations. We found 554 sensitive words and 1,288 deformations covering pinyin, abbreviation, and splitting modes, storing them in a sensitive word table. shows partial examples of identified deformations.

4.3.1 Comparative Analysis of Sensitive Word Recognition Algorithms

We calculated text sensitivity to provide reference for platform processing, setting thresholds and . In Experiment 1, we compared our RSWDT algorithm with the improved deterministic finite automaton algorithm (ST-DFA) and sensitive information decision tree filtering algorithm (SWDT-IFA) across varying text counts and lengths.

Using 1,500 texts containing deformations randomly divided into five groups (100, 200, 300, 400, 500 texts), recall and precision comparisons are shown in [Figure 3: see original paper]. With 100 texts, all algorithms showed lower recall and precision due to insufficient data. As data volume increased, RSWDT stabilized, achieving 95% recall and precision at 500 texts. ST-DFA showed low precision with decreasing trends, while SWDT-IFA exhibited large fluctuations. RSWDT outperformed both algorithms because it effectively recognizes deformations, whereas ST-DFA and SWDT-IFA only handle simple pinyin-based deformations.

For text length impact, we divided Weibo texts (max 140 characters) into five categories: micro (0-28), short (29-56), small (57-84), medium (85-112), and large (113-140). Testing 500 texts per category, results are shown in [Figure 4: see original paper]. For texts under 28 characters, all algorithms performed well. As length increased, RSWDT's recall and precision slowly decreased and stabilized, reaching 95% recall at 112-140 characters. ST-DFA and SWDT-IFA showed fluctuating and significantly decreasing recall. RSWDT's precision stabilized at 95%, while the other algorithms showed lower precision with clear downward trends.

4.3.2 Effectiveness Comparison of Three Deformation Types

Using the same 1,500 texts divided into five groups, we validated RSWDT's effectiveness across the three deformation types. Results are shown in [Figure 5: see original paper]. At 100 texts, all deformation types showed lower recall and precision, but stabilized as data increased. At 500 texts, pinyin mode achieved the highest recall (93%) and precision (94%). For recall: pinyin mode > abbreviation mode > splitting mode, primarily because Chinese character structure is complex and manual splitting analysis may be incomplete, while abbreviation patterns are diverse. For precision: splitting mode > pinyin mode > abbreviation mode, because splitting structures are fixed and errors are rare when recognized, whereas pinyin mode suffers from easily confused pinyin interference.

4.3.3 Impact of Sensitivity Threshold Settings on Filtering Effectiveness

To validate filtering feasibility, we selected 2,132 texts, randomly divided into four samples of 533 texts each. One hundred people identified sensitive words and judged text sensitivity, classifying results into three categories: high sensitivity (A), medium sensitivity (B), and no processing needed (C). Average sensitive word counts were 7 for A, 4 for B, and 2 for C, confirming that people generally judge sensitivity based on word frequency.

Using our sensitivity calculation method with thresholds and , we set high threshold at 0.8 and tested low thresholds at 0.3, 0.4, and 0.5. Results compared with manual judgments are shown in [FIGURE:6(a)], with maximum overlap at . Setting at 0.3 and testing high thresholds at 0.6, 0.7, and 0.8, results in [FIGURE:6(b)] show maximum overlap at . Therefore, and produce results closest to human judgment.

Experiments demonstrate that RSWDT achieves high accuracy for Chinese sensitive word deformation recognition. The sensitivity calculation comprehensively reflects text sensitivity, reducing manual workload and providing intuitive, reliable basis for processing sensitive information, enabling effective automatic filtering.

---

5 Conclusion

The decision tree-based sensitive word deformation algorithm effectively recognizes three deformation types: pinyin mode, abbreviation mode, and splitting mode. The improved multi-factor model calculates text sensitivity for automatic filtering. Our algorithm significantly improves accuracy and efficiency for sensitive information recognition and filtering, particularly for deformations, with experimental results approaching human judgment. However, this research lacks semantic analysis between words, requiring manual judgment of text orientation. When processing large volumes of sensitive texts, the workload becomes substantial. Therefore, character relationships and semantics represent important directions for future work.

References

[8] Fu Cong, Yu Dunhui, Zhang Lingli. Study on the identification method for the change form of Chinese sensitive words [J]. Application Research of Computers, 2019, 36(4).

[9] Chang J S, Teng Weilun. Mining atomic Chinese abbreviation pairs: a probabilistic model for single character word recovery [J]. Language Resources and Evaluation, 2007, 40(3/4): 367–374.

[10] Chinese Character Reform Commission. Scheme of the Chinese phonetic alphabet [S]. 1967.

[11] Yin Zhiping. Methods and principles for the construction of abbreviations [J]. Language Teaching and Linguistic Studies, 1999(2): 73-82.

[12] Zhu Wenxuan. Automatic extraction technology of blog text content sensitive information [D]. Shanghai: Shanghai Jiaotong University, 2008.

[13] Nghia L T, Huy A Q, Ngoc A N. Application of fuzzy-analytic hierarchy process algorithm and fuzzy load profile for load shedding in power systems [J]. International Journal of Electrical Power & Energy Systems, 2016, 77: 178-184.

[14] Lan S, Zhang H, Zhong R Y, et al. A customer satisfaction evaluation model for logistics services using fuzzy analytic hierarchy process [J]. Industrial Management & Data Systems, 2016, 116(5): 1024-1042.

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