Regression Analysis-Based Topic Crawler for Online Terrorist Information: Postprint

Authors: Huang Wei, Zhang Zhancheng, Zhu Bin, Yuefeng Li, Lu Wei

Date: 2023-08-26T00:00:00+00:00

Abstract

[Purpose/Significance] To address the current problems of difficulty and low efficiency in collecting cyber terrorism information from open-source network information, this paper proposes a regression analysis method that integrates semantic relevance and webpage importance to improve collection efficiency. [Method/Process] By analyzing and comparing the characteristics of focused crawlers, combined with the features of cyber terrorism information, the advantages applicable to terrorism information collection in the PageRank and TF-IDF algorithms are identified. Through regression analysis, the terrorism information collection strategy undergoes relevance prediction, and the prediction results are used to feedback-regulate the collection process. [Results/Conclusion] Cyber terrorism information collection must balance both quantity and quality. Based on improvements to traditional focused crawler algorithms, an optimized crawler algorithm specifically for open-source cyber terrorism information collection is proposed, which can enhance information collection efficiency.

Full Text

A Network Counter-Terrorism Information Crawler Based on Regression Analysis

Huang Wei[1], Zhang Zhancheng[1], Zhu Bin[1], Li Yuefeng[1], Lu Wei[2] [1] School of Economics and Management, Hubei University of Technology, Wuhan 430068 [2] Wuhan East Lake High-tech Development Zone Power Company, State Grid Corporation of China, Wuhan 430073

Abstract

[Purpose/Significance] Aiming to address the difficulty and low efficiency of acquiring terrorist information from open-source networks, this paper proposes

a method based on regression analysis to improve the collection efficiency of network terrorist information by combining semantic relevance and webpage importance.

[**Method/Process**] By analyzing and comparing the characteristics of focused crawlers and integrating them with the features of network terrorist information, we identified the applicable advantages of the PageRank algorithm and TF-IDF algorithm for terrorist information collection. Combined with regression analysis, we performed relevance prediction on terrorist information collection strategies and used the prediction results to feedback and regulate the information collection process.

[**Result/Conclusion**] Network terrorist information collection must consider both quantity and quality. Based on improvements to traditional focused crawler algorithms, this paper proposes an optimized crawler algorithm specifically for open-source network terrorist information collection, which can improve information collection efficiency.

Keywords: focused crawler; regression analysis; network counter-terrorism; semantic similarity

---

Introduction

The concept of combining cyberspace with terrorism was first proposed in 1997 by B. Coins, a senior researcher in intelligence and security in California, USA, who considered cyber terrorism as the product of the combination of network and terrorism [1]. In today's big data era, where the Internet is the primary source of information for people, network terrorism continues to develop by exploiting the convenience of the Internet age. Terrorists use the Internet to publish network terrorist information, creating social panic on one hand, and on the other hand, using the Internet to disseminate organizational and planning information for terrorist activities, providing prerequisite information conditions for terrorist attacks. Li Benxian et al. believe that combating terrorism from the Internet perspective is an important direction for network counter-terrorism research [2]. Li Ou points out that although network counter-terrorism has gained worldwide attention, terrorist organizations are becoming increasingly flexible in tactics, more professional in operational capabilities, and more information-based in organization [6]. Faced with these new characteristics, the demand of cybersecurity workers and counter-terrorism departments to collect terrorist information data in the vast ocean of data is growing stronger. The collection of network counter-terrorism data is the foundation of network counter-terrorism, and customized vertical search engine technology is the key to network terrorist information collection [7]. Based on a review of relevant literature in the counter-terrorism field, this paper summarizes the characteristics of network terrorist information from both qualitative and quantitative perspectives into two major categories: digital features and essential features, as shown in Fig-

ure 1 [Figure 1: see original paper], and proposes a specific focused crawler for network counter-terrorism data collection through improvements to existing focused crawler algorithms.

The characteristics of network technology—its vulnerability, the concealment of network activities, the richness and openness of network resources, and the popularity and widespread nature of networks—are exploited by terrorists, making it their "second battlefield" for conducting terrorist activities [3]. The vertical characteristics of terrorist information are elements used for quantitative processing and discrimination of the credibility of network terrorist information, called digital features, which can be calculated algorithmically. The horizontal characteristics are abstract features used to describe network terrorist information, which can only be determined manually based on experience, called essential features. This paper proposes a regression analysis model for the digital features of network terrorist information to gradually adapt various terror-related factors to a curve, comprehensively judging the various features of network terrorist information, thereby improving the collection precision of focused crawlers for network terrorist information.

With the rapid development of the information age, current network terrorism is becoming globalized, with international terrorist organizations closely connected to regional terrorist organizations [4]. At the "Network Counter-Terrorism Forum" held during the third World Internet Conference in Wuzhen in 2016, it was proposed to strengthen international cooperation and jointly combat network terrorism [5]. Although network counter-terrorism has gained worldwide attention, terrorist organizations are becoming increasingly flexible in tactics, more professional in operational capabilities, and more information-based in organization [6]. Faced with these new characteristics, the demand of cybersecurity workers and counter-terrorism departments to collect terrorist information data in the vast ocean of data is growing stronger. The collection of network counter-terrorism data is the foundation of network counter-terrorism, and customized vertical search engine technology is the key to network terrorist information collection [7].

This work is supported by the National Natural Science Foundation of China projects "Research on Multi-kernel Methods for Real-time Active Perception of Network Public Opinion Events in Microblog Environment" (Project No.: 71303075) and "Research on Unsupervised Text Classification Method Based on Feature Ontology Learning in Big Data Environment" (Project No.: 71571064).

Author Introductions: Huang Wei (ORCID: 0000-0002-5804-9371), Professor, Ph.D., Master's Supervisor, E-mail: tonyhw@163.com; Zhang Zhancheng (ORCID: 0000-0002-7533-4764), Master's Student; Zhu Bin (ORCID: 0000-0002-1073-0379), Undergraduate; Li Yuefeng (ORCID: 0000-0001-5173-9575), Professor, Ph.D.; Lu Wei (ORCID: 0000-0001-6270-8846), Assistant Engineer.

Received: 2017-08-21 **Revised:** 2017-11-16 **Pages:** 121-129 **Responsible Editor:** Wang Shanjun

2. Focused Crawler and Algorithm Review

A focused crawler is an intelligent agent specifically designed to collect specific topic document information on the Internet, capable of automatically searching and crawling on the Internet and returning the collected topic information to the server [8]. The workflow diagram of a focused crawler is shown in Figure 2 [Figure 2: see original paper].

The key to focused crawler collection of network information lies in how to accurately analyze the relevance between webpage content and topics, and what collection strategies should be used to make the collection process more efficient and timely. Research on focused crawler algorithms is divided into two categories: one based on link structure [9], and the other based on content evaluation [10]. The representative algorithm of the former is the PageRank algorithm, while the latter includes the TF-IDF algorithm.

In the research of network focused crawler search strategies, in addition to the above-mentioned PageRank-based link evaluation, another category of focused crawler algorithms evaluates webpage content, with the representative algorithm being TF-IDF. Lu Yonghe et al. combined TW with TF-IDF as a new feature weight algorithm [19]. Wang Jingzhong et al. combined regular expressions and semantic analysis technology to improve the TF-IDF algorithm [20]. In evaluating webpage content and topics, the improved TF-IDF fully utilizes tags, anchor text, and stop word filtering in original webpages to adjust the weight values assigned to keywords in webpages, thereby more accurately determining the relevance between webpages and topics. However, network terrorist information is characterized by vague topics, scattered and disorganized content, and high information uncertainty. Current improved TF-IDF algorithms cannot effectively utilize the interrelationships between webpage links, and thus cannot effectively make correlation judgments on terror-topic-related webpages or collect them according to the relevance of network terrorist information.

In webpage evaluation algorithms based on link structure, the representative is the PageRank algorithm. Yang Bin et al. proposed a concept-based weighted PageRank improvement algorithm [11]. Lin Hong et al. proposed a PageRank improvement algorithm based on the unequal probability of users clicking various links within webpages to avoid topic drift [12]. He Ming et al. proposed a semantically similar PageRank improvement algorithm [13]. Wang Zhongfei et al. proposed a PageRank improvement algorithm based on anchor text similarity [14]. Wang Jianxiong improved the traditional PageRank algorithm by calculating the similarity between hyperlinks and domain vectors to suppress topic drift, introducing time factors and internal/external differentiation factors to increase the weight of topic-related URLs, thereby improving information collection efficiency [15]. Wang Chong et al. proposed a PageRank improvement algorithm based on user interest and topic relevance [16]. The PageRank algorithm is the

core of focused crawler algorithms and is applied in vertical search engines for user-specific topics, capable of better determining collection strategies. However, in the PageRank algorithm and its current improvements, adjustments to link importance in the network are based on network information accessed by users for specific topics [17]. For example, Yang Bin and Lin Hong adjusted the weight of URL queues in webpages from the perspective of user browsing, and J.M. Maestre et al. applied PageRank to joint control of target groups [18]. However, network terrorist information pages are not frequently visited by general users, so webpage importance cannot be simply determined by link clicks, visits, or the number of incoming/outgoing links. In fact, webpages with more links are more likely to be unrelated to network terrorism because network terrorist information is highly covert. It differs from webpages searched and ranked by popularity in traditional search engines, so webpage popularity cannot be used as a key indicator for judging terror-themed webpages.

3. System Framework Design

Faced with the extreme problems existing in traditional focused crawler algorithms, this paper proposes a regression analysis prediction model by combining improved PageRank algorithm and improved TF-IDF algorithm. Since judging whether network information is terror-related cannot be determined by one or several topic keywords alone, nor can it be determined solely by webpage link numbers and click volumes. Instead, these related factors must be organically combined through analysis to improve the quality and efficiency of network terrorist information collection. This paper adopts the method from the ontology-based network group event topic discovery model [21] to establish our own network ontology terrorist information database, thereby fully utilizing already collected network terrorist information data to provide data support for regression analysis of network terrorist information.

Traditional focused crawler systems need to provide a certain number of initial URLs and topic words in advance. Initial URLs are used to parse and obtain other webpage pages and extract secondary or multi-level URLs in webpages to form an initial URL queue. The above process is cycled to obtain a URL library to be crawled, while topic words are used to judge the relevance between crawled text and topics. However, selecting high-quality topic keywords that can comprehensively describe network terrorist information is very difficult. This paper proposes using regression analysis prediction methods to solve the dilemma of collecting network terrorist information. The specific approach is as follows: 1) By providing several network terrorist information pages, the regression analysis prediction module analyzes and filters out terror topic keywords, stores the keywords and page link numbers, establishes a terrorist information word list, and uses the keywords and link numbers in the list as labels for each terror keyword. 2) Through regression of terror keywords and terror-related URLs, set regression analysis offset to control the number of keywords for each network

terrorist information collection, providing keywords and initial URL queues for the focused crawler. 3) Parse initial URLs to corresponding webpages, collect webpage content by the collection module, and analyze the number of webpage topic keywords and URL links and other webpage information by the analysis module. 4) Through regression of terror webpage information similarity, parse collected webpages according to certain rules, combine various factors for judging terror information into a curve through regression functions, and select the priority order of network terrorist information collection based on the goodness-of-fit with authoritative terror information webpages. 5) Through regression analysis, store keywords, webpage link numbers, and other factors from webpages with high goodness-of-fit into the network ontology terrorist information database. Terror information webpages with lower goodness-of-fit but within the set minimum threshold are added to the network terrorist information database through manual review to provide reference for the next regression analysis collection. The regression analysis focused crawler structure diagram is shown in Figure 3 [Figure 3: see original paper].

---

4. Regression Analysis

4.1 Improved PageRank Algorithm In the process of network counter-terrorism information collection, quantifiable features include relevance and comprehensiveness, which can be measured by the improved PageRank algorithm. As an important algorithm for Google to evaluate webpages, the most significant feature of the PageRank algorithm is its objective reflection of webpage importance in Web links [22]. The principle lies in webpages "voting" through links, obtaining the situation of webpage inbound and outbound links, measuring their "link popularity." That is, for a certain webpage node, the greater the number of inbound links, the greater the importance of the webpage. Webpage importance is quantified through PR values, with PR values ranging from [0,10], where larger PR values indicate greater popularity [23].

For a certain terror-related webpage M, the PageRank algorithm follows two assumptions: (1) Quantity assumption: If terror webpage M has more inbound terror webpage links and strong relevance, it indicates greater influence and higher importance. (2) Quality assumption: The more important the network terrorist information webpage, the higher the terror weight it receives. If other webpages link to it, its importance is relatively higher.

Assuming webpages I_1, I_2, I_3, …, I have inbound links to webpage M, and the total number of links for the i-th webpage is L(i), the PR value of webpage M is calculated based on the weight distribution of links from webpages, as shown in Equation (1):

$$PR(M) = \frac{PR(I_1)}{L(I_1)} + \frac{PR(I_2)}{L(I_2)} + \frac{PR(I_3)}{L(I_3)} + \frac{PR(I_n)}{L(I_n)}$$

Since there exist webpages that do not link to any other links, also called "dead links," which would cause errors in the formula, a damping factor (d) is added to correct the formula, with d generally taking the value 0.85 [24]. The random surfer model also confirms the usefulness of the damping factor, indicating that a webpage surfer will not keep clicking a link but will randomly jump to other webpages, ensuring average weight distribution for each outbound link. The modified formula is shown in Equation (2):

$$PR(M) = (1-d) + d\left(\frac{PR(I_1)}{L(I_1)} + \frac{PR(I_2)}{L(I_2)} + \frac{PR(I_3)}{L(I_3)} + \frac{PR(I_n)}{L(I_n)}\right)$$

4.2 Improved TF-IDF Semantic analysis is an important factor in topic relevance analysis. When analyzing webpage importance alone, whether topic words match webpage content is not considered, which may lead to "topic drift." This paper adopts the TF-IDF semantic analysis algorithm to analyze feature words in webpage text content and calculate webpage topic relevance based on the distribution of network terror feature words. The TF-IDF algorithm had certain defects at its inception. Z.H. Deng et al. proposed alternative CRF (category relevance factors) [25], and Zhao Xiaohua et al. proposed using feature selection to correct function weights. The TF-IDF-CHI algorithm improved the initial algorithm to some extent.

Based on previous research on network terrorist information, a relatively reasonable tag weight function m(i) was obtained, as shown in Equation (4) [27]:

$$m(i) = \begin{cases} 10, & \text{title} \\ 8, & \text{meta} \\ 6, & H, a \end{cases}$$

Lu Yonghe et al. [19] addressed the distribution of feature words within and between text categories, proposing that for feature words and document categories, the number of documents C that do not contain the feature word but belong to the category, and the number of documents B that contain the feature word but do not belong to the category, plus the commonly used feature selection evaluation function chi-square value CHI. It was concluded that the smaller the C and B, the more uniform and dispersed the intra-class distribution, and the more highly concentrated the inter-class distribution, the greater the feature weight value of this feature term. This weight is derived as shown in Equation (5):

$$TW(i) = \log CHI_i \times \log(B_i \times C_i)$$

where i is a feature term of terror information, CHI_i is the CHI value of this terror feature term, B_i is the number of terror-related texts that contain i

but do not belong to this category, and C_i is the number of terror-related texts that do not contain i but belong to this category. Due to the ambiguity and concealment of network terrorist information, for some terror information feature terms, B_i × C_i may be 0. Therefore, a relatively small constant λ is added to B_i × C_i, resulting in the weight formula shown in Equation (6):

$$TW(i) = \log CHI_i \times |\log(B_i \times C_i + \lambda)|$$

Wang Jingzhong et al. [20] pointed out in their improved TF_{IDF} algorithm that assigning weights to keywords within feature words is very important because keywords that contribute in feature words may contain empty words like "的" (de) and "呢" (ne), making weight assignment meaningless. Therefore, empty words need to be removed, and different weights should be assigned to keywords and non-keywords. Feature words occupy a high proportion in anchor text, title, and meta tags [27]. Considering that contribution degree may shift when these tags appear alone, an evaluation weighted accumulation calculation method is adopted to derive the weight formula shown in Equation (3):

$$T_{wf}(k) = \frac{\sum_{i=0}^{n} m(i)}{\sum_{j=1}^{N} m(j)}$$

where m(i) is the weight value of the i-th tag, T_{wf}(k) refers to the average cumulative weight value of the k-th word, *{i=0}^{n} m(i) represents the cumulative weight of the tag, and* {j=1}^{N} m(j) represents the total weight sum of the above tags contained in the entire page.

Finally, based on the contribution degree of keywords in feature words and the distribution relationship between document categories, a comprehensive TF-IDF formula for calculating semantic relevance is derived as shown in Equation (7) [16]:

$$W_{ik} = TW(t) \times T_{wf}(k) \times \log(tf_{ik}) \times idf_{ik} = TW(t) \times T_{wf}(k) \times \log(\varepsilon + N/n_k)$$

where W_{ik} represents the weight of the k-th terror-related document containing the i-th terror feature word, TW(t) represents the term correction weight for the distribution of terror feature words across terror document categories, T_{wf}(k) represents the keyword weight of the k-th document, N represents the total number of documents, and n_k represents the number of documents containing the feature word. Due to varying text lengths of network terrorist information, longer terror-related texts may have larger weights for terror keywords. To solve this problem, a logarithmic approach log(tf_{ik}) × idf_{ik} = log(+ N/n_k) is used for standardization to reduce the impact of excessive weights caused by overly long texts.

4.3 Regression of Terror Keywords and Terrorism-related URLs The working principle of a focused crawler is to parse the seed URL collection, extract the corresponding URL queue, download the pages corresponding to the URL queue, and parse the series of URLs contained in the pages to expand the URL collection. Each crawler cycle forms the URL collection for the next collection, so URLs have a significant impact on the efficiency and quality of focused crawler information collection [28].

The core of regression analysis is that when extending from initial URLs to other URLs, the relevance between the content of webpages corresponding to other URLs and the topic is linearly related to the relevance of the content corresponding to the initial URLs to the topic. Simply put, webpages with higher topic relevance contain Web links whose content is also more relevant to the topic. Based on this idea, this paper first parses initial URLs, extracts URLs and vocabulary in webpages, and establishes a unidirectional index of URLs and a multiple index of keywords. After removing duplicate keywords with high repetition rates but obvious deviation from the topic, such as empty words, a topic vocabulary list is created. Then, the URL importance PR(M) from the PageRank algorithm is introduced, and the filtered keyword list is labeled and correlated with URL importance PR(M), as shown in Figure 4 [Figure 4: see original paper]:

$$P_1 = \{[PR(M)]_1 + [PR(M)]_3 + [PR(M)]_j + ...\} \cdot \rho + c \cdot W_{ik}$$

Equation (8) represents the importance of each keyword among numerous URLs in the collection process. P_1 represents the importance of the first topic keyword, [PR(M)]*j is the importance corresponding to the keyword in webpage j, W*{ik} is the cumulative weight of the keyword in the TF-IDF algorithm, and and c are parameters used to adjust the relationship between the two variables.

Through the above method, the factor of URL importance or popularity is introduced into the evaluation of webpage relevance to topics by topic keywords, assigning certain weights to keywords that appear infrequently in crawled content but have high topic relevance, and keywords that appear frequently but have low topic relevance. Different importance levels of keywords are treated differently, thereby improving the accuracy of keyword assessment of webpage content.

The importance of all keywords contained in a webpage is recorded as X, X = _{i=1} p_i, and the importance of the webpage is recorded as Y, Y = PR(M)^n, as shown in Equation (9):

$$Y = A + BX + \epsilon$$

$$B = \underline{\sum xy - n \cdot \sum x \sum y / [\sum x^2 - n \cdot (\sum x)^2]}$$

$$A = \frac{\sum y - B \cdot \sum x}{n}$$

Correlation regression analysis is performed, which is the regression analysis 2 module diagram. The regression line obtained from regression analysis is shown in Figure 5 [Figure 5: see original paper]:

Two straight lines are drawn as shown in Equations (10) and (11):

$$Y_1 = (A + v)x + B + \epsilon$$

$$Y_2 = (A - v)x + B + \epsilon$$

A and B are parameters to be determined, where A is the intercept of the regression line and B represents the slope of the regression line, showing the average change in Y for each unit change in X. represents random error with user satisfaction as a reference factor. Q represents the minimum topic relevance, and V represents the tolerance value. Q and V are two thresholds that need to be predefined for the focused crawler to adjust and control the range of crawled URLs. Q is a parameter ensuring the relevance between page content and topic, and V is the maximum allowable irrelevance between URLs and topic. The crawler only needs to crawl the area enclosed by Q and Y_1 and Y_2, eliminating URLs outside this area, and using the remaining URLs as the URL collection for the next cycle. As shown in the terror keyword and terror-related URL regression module in Figure 3, when V = 0.25, almost all terror-related URLs are included. This approach greatly reduces the number of URLs, improves the quality of URL queues, reduces the workload of the focused crawler, and improves system operational efficiency.

4.4 Regression of Terror Web Page Information Similarity Network terrorist information has characteristics such as timeliness and dispersion. Based on these potential characteristics that may exist in network terror-related information, this paper combines various factors used to identify terror information in the same webpage, such as keywords, link numbers, occurrence time, and visit volume, into a logical curve, as shown in Equation (12):

$$Z = \beta_0 + \beta_1\phi_1 + \beta_2\phi_2 + ... + \beta_k\phi_k$$

where β_i are called regression parameters and ϕ_i are various terror influence factors. At the beginning of regression analysis, the values of ϕ_i are set manually based on experience or expectations, similar to the weights in the HITS algorithm. As crawled content continues to increase, network terror information crawled from the Internet is deduplicated and denoised, and this category

of network terror information webpages scores ϕ_i, adjusting ϕ_i according to score magnitude.

Equation (12) is transformed through a regression function as shown in Equation (13):

$$f(z) = \frac{e^{-z} + 1}{}$$

The function graph is shown in Figure 6 [Figure 6: see original paper]:

The characteristic of this algorithm's regression model is that the variable range is from -∞ to +∞, but the value range is between (0-1), thus transforming multiple terror factors into a probability to judge the relevance of network terrorist information. (β_0, β_1, β_2, ..., β_k) are grouped according to the P_n of topic keywords assigned with webpage link numbers. Each time a new webpage is obtained: 1) Words and links in the webpage are extracted according to webpage modules. 2) The P value in each webpage module is calculated and matched with the corresponding P value under the main keyword in the database. 3) The group (β_0, β_1, β_2, ..., β_k) under this P value label in the database is used as parameters for key information in the webpage to calculate the Z value, and the value on (0-1) corresponding to f(z) is used to judge the degree of terror information in the webpage.

For cases where key information ϕ_k may be missing in a webpage module, first, the original ϕ_k in the database is supplemented during calculation, and then the logarithm of the missing terror factor and log(v§) (logarithm of error value and tolerance value) is subtracted to reduce the impact on regression analysis discrimination of network terrorist information credibility. Finally, the formula for calculating z becomes Equation (15):

$$z = \beta_0 + \beta_1\phi_1 + \beta_2\phi_2 + ... + \beta_k\phi_k - \sum_{i=1} \beta_j\phi_j \log v\S$$

5. Experimental Results Analysis and Evaluation

To verify the effectiveness of the improved focused crawler algorithm with regression analysis in crawling open-source network terror information, the experiment compares the improved semantic similarity PageRank algorithm, improved TF-IDF algorithm, and regression analysis algorithm to prove the superiority of the focused crawler with regression analysis in collecting network terrorist information.

5.1 Experimental Design To verify the advantages of the above improved algorithm, 2,000 texts were collected as the experimental data source. Among

them, 1,000 texts were collected through the network terrorist information collection system and partially manually screened to confirm terror information texts. The other 1,000 texts were ordinary network text information collected by a general crawler, but terror keywords from the network terror lexicon were randomly inserted at random positions in each ordinary text to constitute experimental samples of terror information texts.

5.2 Establishment of Counter-Terrorism Lexicon The establishment of a counter-terrorism lexicon is the most important aspect of regression analysis. Only by establishing a correspondence between terrorism-related vocabulary in the lexicon and links can new webpages be further related to vocabulary and labels in the lexicon, and terror information regression analysis be carried out to judge the relevance between webpages and terror topics. After manual analysis of 10,000 pieces of network terrorist information data crawled from the Internet, a 100-word terror topic information lexicon table was established, with a partial table shown in Table 1 . The first letters A, B, C, etc., in the table are categories of terror information words. Lexicon categories are classified according to word attributes, such as location nouns, event nouns, terrorist organization codes, code words, etc. The numbers after letters represent their serial numbers. The more digits the serial number has, the higher the importance of the lexicon, and the corresponding weight is relatively higher.

The process of lexicon establishment is shown in Figure 7 [Figure 7: see original paper]:

Experimental Process: 1. Perform word segmentation and deduplication on the terror information lexicon through regression analysis. 2. Calculate semantic similarity based on the lexicon and make judgments. Add information of identical words to the original lexicon, and perform semantic similarity calculation on the new terror information lexicon. 3. Calculate relevance with existing words in the lexicon, compare the calculated results with the defined threshold of (0, 0.85). Relevance in (0.85, 1) indicates high semantic relevance with words in the lexicon, and they are directly added to the lexicon. If relevance is in (0, 0.85), indicating not very high relevance with words in the lexicon, manual judgment is performed, and terror information words after manual judgment and analysis are added to the final terror information lexicon.

Through collection and organization of network terrorist information, it was found that the generation, development, and evolution of network terrorist information follow certain patterns. Terror information hidden in networks is distributed in a network pattern according to its keywords and links as labels. Therefore, this paper proposes establishing a terror information word list. The establishment of the word list is similar to the network diagram of CNKI person relationship diagrams, dividing different lexicons into different levels according to the number of lexicons, link numbers, and linked numbers contained in important texts, as shown in Figure 8 [Figure 8: see original paper]:

In the figure, a circle represents a keyword, the number in the circle represents the number of historical network terrorist information documents corresponding to that keyword, and the lines represent relationships between terror information keywords. The number of links is reflected in Table 1. The establishment of the network terrorist information word list is based on the relevance and dispersion shown in the digital features of network terrorist information in Figure 1, establishing correspondences between topic words, URLs, and ontology content of network terrorist information. This can not only provide a basis for subsequent analysis of network terrorist information characteristics and network counter-terrorism measures but also provide data reference and clue guidance for the next network terrorist information collection, improving the efficiency of the next network terrorist information collection.

5.3 Recall and Precision The commonly used performance indicators for focused network crawlers are precision and recall. The precision formula is P = K/N, where K is the number of pages crawled related to the topic, and N is the total number of pages crawled. Recall, also called recall rate, is calculated as: R = K/R, where R is the total number of pages related to the topic existing on the network. To ensure effective use of the 2,000 test data sources, the experiment selected representative URLs with large numbers of outbound links from various terror text information and non-terror text information as initial URLs for collection. Keywords were those already established in the network counter-terrorism lexicon. Finally, the 2,000 texts containing terror information were crawled by three crawler algorithms and compared to obtain experimental results, as shown in Table 2 :

In Table 2, the relevance threshold for PageRank links is 0.3. In $\Delta f(z)V(0.2)$, $\Delta f(z)$ represents the absolute value between the actual f(z) value of terror-related network terrorist information and the standard $f(z)_0$ of network terrorist information in the regression of terror webpage information similarity, i.e., $\Delta f(z) = |f(z) - f(z)_0|$. V represents the tolerance value for regression prediction of URL queues in the regression of terror keywords and terror-related URLs. Here, the value of V is constantly 0.25. The selection of V value depends on the quality of initial URLs and the number of terror-related lexicons in the lexicon corresponding to initial URLs. As can be seen from the table, the selection of $\Delta f(z)V(0.2)$ is very important. When its value is around 0.2, that is, when V = 0.25 and $\Delta f(z) = 0.8$, both recall and precision perform well. Compared with general crawlers and crawlers based on the PageRank algorithm, the crawler algorithm using regression analysis can greatly improve the accuracy of collecting similar network terrorist information, avoid blindly collecting in massive network data, and improve information collection efficiency. The line chart corresponding to Table 2 is shown in Figure 9 [Figure 9: see original paper]:

5.4 Discussion and Analysis From Figure 9, it can be seen that: 1) The regression analysis algorithm proposed in this paper based on PageRank algorithm and TF-IDF algorithm can well improve the crawling efficiency of web-

pages with characteristic terror information on the network, and precision is indeed greatly improved. 2) In the process of using regression analysis crawler, it was found that changing $\Delta f(z)V$ has a significant impact on results. As seen from regression analysis crawler 1, increasing $\Delta f(z)$ can make newly collected network terrorist information have higher similarity and relevance with old network terrorist information, but it will reduce the quantity of network terrorist information collected. 3) The line segment from regression analysis 2 to regression analysis 3 shows that appropriately increasing the $\Delta f(z)$ value can improve recall while precision decreases relatively little. This paper calls this period the equilibrium period. Generally, to avoid missing any information, the value of $\Delta f(z)$ should be selected at the end of the equilibrium period to achieve the purpose of crawling all text information with the same characteristics as terror information texts. After adjustment, $\Delta f(z)$ can fully reflect the differences and characteristics between terror information texts, compare similarity between different types of terror information texts, enable focused crawlers to collect network terrorist information targeted, and improve the efficiency of network terrorist information text mining.

With the advent of the network information age, network information has exploded. Terrorists take advantage of the Internet to disseminate terrorist organization information or terror public opinion information, posing a huge threat to the nation and society. How to collect useful information related to terrorists from massive open-source network information, analyze this information, and prevent and stop network terror events is a problem to be solved. The crawler system using regression analysis proposed in this paper can crawl valuable network terrorist information from massive network information, providing intelligence and basis for counter-terrorism work. This algorithm is a comprehensive application based on existing PageRank and HITS core algorithms and focused crawler strategies, transforming unrelated algorithm variables into probabilities through regression analysis prediction, combining several factors to find connections between terror information, and setting thresholds of certain sizes to screen out network terrorist information, achieving the purpose of improving information collection efficiency. The deficiency of this algorithm lies in the simultaneous use of two algorithms and the combination with regression analysis algorithm, making the algorithm overly complex, slow in actual execution speed, and requiring high hardware performance. How to optimize the algorithm, improve operational efficiency, and information matching accuracy is the difficulty we need to overcome next.

References

[1] C.T. Boucher K. The diffuse border: intelligence-sharing, control and confinement along Canada's smart border [J]. Surveillance & society, 2008, 5(2): 142-165.

[2] Li Benxian, Jiang Chengjun, Fang Jinqing. Challenges and opportunities faced by network science in counter-terrorism research [J]. Complex Systems and Complexity Science, 2014, 11(1): 60-66.

[3] Li Ou. Network counter-terrorism and countermeasures [J]. Journal of Jiangxi Police College, 2006(3): 92-96.

[4] Wang Yong, Mei Jianming. Characteristics, challenges, and coping strategies of current counter-terrorism struggle [J]. Journal of Chinese People's Public Security University (Social Sciences Edition), 2016, 32(1): 19-23.

[5] Huang Wei, Yu Hui, Li Yuefeng. Research on the construction of network counter-terrorism knowledge base [J]. Journal of Intelligence, 2017, 36(5): 168-174.

[6] Liu Jiong. Research on the prevention and control of violent and terror audio-video dissemination in the network era [J]. Journal of Chinese People's Public Security University (Social Sciences Edition), 2015, 31(1): 1-9.

[7] Huang Wei, Yu Hui, Li Yuefeng. Current status, problems, and prospects of domestic network counter-terrorism research [J]. New Technology of Library and Information Service, 2016(11): 1-10.

[8] Chakrabarti S, Berg MVD, Dom B. Focused crawling: a new approach to topic-specific resource discovery [J]. Computer networks, 2000, 31(11/16): 1623-1640.

[9] Heydon A, Najork M. Mercator: a scalable, extensible Web crawler [J]. World Wide Web: Internet & Web Information Systems, 1999, 2(4): 219-229.

[10] Avraam I, Anagnostopoulos I. A comparison over focused Web crawling strategies [C]//Panhellenic conference on Informatics. Kastoria: IEEE Computer Society, 2011: 245-249.

[11] Yang Bin, Kang Muning. Concept-based weighted PageRank improvement algorithm [J]. Journal of Intelligence, 2006(11): 70-72.

[12] Lin Hong, Liu Peng, Li Jingjing, Long Zhenhai. Probability-based PageRank improvement algorithm [J]. Journal of Wuhan University of Technology, 2009(3): 81-83.

[13] He Ming, Zhou Jun, Li Shuyou. Semantically similar PageRank improvement algorithm [J]. Computer Engineering and Applications, 2009(27): 140-142.

[14] Wang Zhongfei, Wang Biao. PageRank improvement algorithm based on anchor text similarity [J]. Computer Engineering, 2010(24): 258-260.

[15] Wang Jianxiong. PageRank improvement algorithm based on special topics [J]. Library and Information Service, 2012, 56(21): 114-118.

[16] Wang Chong, Ji Xianhui. Research on PageRank algorithm improvement based on user interest and topic relevance [J]. Computer Science, 2016, 43(3): 275-278.

[17] Wang Deguang, Zhou Zhigang, Liang Xu. Analysis and improvement of PageRank algorithm [J]. Computer Engineering, 2010, 36(22): 291-293.

[18] Maestre JM, Ishii H. A PageRank based coalitional control scheme [J]. International journal of control automation & systems, 2017, 15(5): 1983-1990.

[19] Lu Yonghe, Li Yanfeng. Improved TF-IDF algorithm for text feature term weight calculation [J]. Library and Information Service, 2013, 57(3): 90-95.

[20] Wang Jingzhong, Qiu Tongxiang. Focused crawler based on improved TF-IDF algorithm [J]. Computer Applications, 2015, 35(10): 2901-2904, 2919.

[21] Huang Wei, Cheng Baosheng, Yang Qing. Research on ontology-based network group event topic discovery [J]. Library and Information Service, 2012, 56(20): 47-52, 27.

[22] Li Yiying, Yang Wu, Xie Zhijun. Survey of PageRank algorithm research [J]. Computer Science, 2011, 38(S1): 185-188.

[23] Song Juping, Wang Yongcheng, Yin Zhonghang, et al. Improvement of webpage PageRank algorithm [J]. Journal of Shanghai Jiaotong University, 2003(3): 397-400.

[24] Zhu Haodong, Ding Wenxue, Yang Lizhi, et al. Improved PageRank algorithm based on user behavior and topic similarity in microblog environment [J]. Computer Engineering, 2017, 43(5): 179-184.

[25] Deng ZH, Tang SW, Yang DQ, et al. A linear text classification algorithm based on category relevance factors [C]//International conference on Asian digital libraries: digital libraries: people, knowledge, and technology. New York: Springer-Verlag, 2002: 88-98.

[26] Zhao Xiaohua, Ma Jianfen. Improvement of word weight calculation method in text classification algorithm [J]. Computer Knowledge and Technology, 2009, 5(36): 10626-10628.

[27] Ye YX. New research advances in semantic Web search technologies [J]. Computer Science, 2010, 1(37): 1-5.

[28] Zhang Huan, Liu Naiwen, Duan Huichuan. Research on focused crawler based on T-Graph algorithm [J]. Computer Engineering and Design, 2014, 35(9): 3014-3017, 3028.

Author Contributions

Huang Wei: Proposed the article idea and drafted the initial content; **Zhang Zhancheng:** Wrote the research status and was responsible for algorithm implementation; **Zhu Bin:** Organized and analyzed network terrorist information and provided experimental data; **Li Yuefeng:** Studied the algorithm and optimized the paper; **Lu Wei:** Analyzed experimental data.

A Network Counter-terrorism Information Crawler Based on the Regression Analysis

Huang Wei[1], **Zhang Zhancheng**[1], **Zhu Bin**[1], **Li Yuefeng**[1], **Lu Wei**[2] [1] School of Economics and Management, Hubei University of Technology, Wuhan 430068 [2] Wuhan East Lake High-tech Development Zone Power Company, State Grid Corporation of China, Wuhan 430073

Abstract: [Purpose/significance] Aiming at the problems that getting the terrorist information on the network is difficult and the acquisition efficiency is low from the open source network information, a method based on the regression analysis is proposed to improve the acquisition efficiency of the network terror information by combining the advantages of the semantic relevance and the webpage importance. [Method/process] By analyzing and comparing the characteristics of the theme crawler and combining them with the characteristics of the network terrorist information, the advantages of the PageRank algorithm and the IF-IDF algorithm for the collection of the terrorist information were found out. Combined with terrorism semantic similarity regression analysis, the relevance prediction of the terrorist information was done, which reflected the process of the information collection. [Result/conclusion] Both the quantity and quality of the collection of the network terrorist information should be taken into consideration. Based on the traditional common network crawler algorithm, this paper proposes a crawler optimization algorithm pertinent to the network terrorist information collection, which improves the collection efficiency.

Keywords: theme crawler; regression analysis; network anti-terrorism; semantic similarity

Data and Knowledge Organization, Communities of Practice and Knowledge Operation, Content Management and Knowledge Sharing, Knowledge Creation and Open Innovation, Data Mining and Knowledge Discovery.

Submission Guidelines: 1. Manuscript themes should be knowledge-related, discussing knowledge management, knowledge services, knowledge innovation, and related issues. Articles may focus on theory or emphasize application, technology, methods, models, and best practices. 2. Articles must be substantive, integrate theory with practice, have clear research purposes, appropriate research methods, original academic insights, and have reference, learning, or guiding value for theory or practice. 3. All submissions must undergo similarity detection and peer review, and pass through the editorial department's initial review, re-review, and final review. 4. Article length is not limited, but generally 4,000-20,000 words are appropriate. 5. Authors will be informed of acceptance within 1 month. 6. Manuscripts are mainly published online through our website (www.kmf.ac.cn) and authorized databases. Open access and print-on-demand are implemented.

Please submit to: www.lis.ac.cn, noting "Knowledge Management Forum Submission". **Contact:** 010-82626611-6638 **Contact Person:** Liu Yuanying

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