An improved Simhash algorithm based malicious mirror website detection method

Guangxuan Chen¹, Guangxiao Chen²³*, Di Wu⁴, Qiang Liu¹, Lei Zhang⁵ and Xiaoshi Fan⁵

¹Zhejiang Police College, Hangzhou, Zhejiang, 310053, China
²Wenzhou Public Security Bureau, Wenzhou, Zhejiang, 325027, China
³Universidad Carlos III de Madrid, Getafe, Madrid, 28903, Spain
⁴Hangzhou City People's Procuratorate, Hangzhou, Zhejiang, 310016, China
⁵Joint Services Academy, National Defence University, Beijing, 100858, China

*Corresponding author’s e-mail: chenguangxuan@zjjcxy.cn

Abstract. There are a large number of similar or even identical webpages on the Internet. These webpages will cause unnecessary loss of network resources, including waste of storage space, decreased web search speed, and decreased user experience. And some malicious mirror websites will become tools for criminals to carry out illegal activities such as phishing attacks. In this paper, the authors analyzed the mainstream text similarity detection algorithms and webpage deduplication algorithms, and proposed an improved webpage deduplication algorithm based on Simhash. The algorithm converts the text collection into Simhash fingerprints for storage through mapping, and calculates the similarity of the two fingerprints through Hamming distance, thereby obtaining the similarity of the webpage. Experiments show that the algorithm proposed in this paper has a higher accuracy rate and recall rate, and can be better applied to the identification and detection of malicious mirror websites.

1. Introduction

Currently, there is a huge amount of information on the Internet, and a large amount of data is being generated all the time. Much of the information is similar or even completely repetitive. Some of the web pages with a high degree of similarity on the Internet are completely copied, some have slightly modified the content of the web page, and some are only modified in the format of the web page on different websites. Similarity webpages are generally divided into the following four categories: webpage content and format are completely duplicated; webpage content is the same, but format is different; part of webpage content is duplicated with the same format; part of webpage content is duplicated with different format.

These similar web pages have no meaning, they will cause unnecessary loss of network resources, the efficiency of Internet users’ access to information will decrease, and the burden on Internet search engines will also increase. The main impacts include: (1) Waste of storage space. If the database and some storage devices store a large number of similar web pages, it will cause unnecessary storage space waste. Removing these similar web pages can save storage space to store more valuable web pages. (2) The speed of web search has decreased. The existence of a large number of similar webpages on the Internet will increase the indexing, storage, and search overhead of search engines, and will also affect
the retrieval efficiency of search algorithms. If search engines can avoid these similar or identical webpages when searching webpages, it will be greatly improve the speed of searching. (3) The experience of Internet users has declined. If Internet users search for a large amount of similar or even completely repeated information each time, the experience of Internet users will decline; on the contrary, if the search engine presents users with less similar information and a lot of effective information each time, it will improve Internet users’ experience and also improve the efficiency and quality of users’ access to information.

In addition, there is another situation, namely the mirror website. The mirror website mainly includes two forms: one is a mirrored webpage with completely duplicate content; the other one is similarly mirrored webpage with part of the same content. This kind of similarly mirrored webpage is different from the main/source website in terms of webpage format, webpage title, website signature, etc., but the main content presented is the same. While mirroring websites provide convenience, they also bring many problems and harms.

Malicious mirrors are often used in illegal areas, such as illegal websites and web page content plagiarism. By renting some public IP hosts, malicious webpages, illegal webpages and other illegal content are copied to the mirroring host, thereby bypassing control and allowing illegal content to continue to spread. When the normal webpage is maliciously mirrored, the data of the normal webpage is stolen and the content is synchronized. While website mirroring takes up resources, it also affects the loading speed of normal web pages. If it is not processed in time, it will face the danger of being demoted by search engines, or even not included by search engines, especially for new webpages that have just been launched without weight.

Some criminals often carry out illegal activities such as phishing attacks by illegally mirroring some bank websites and shopping websites. Therefore, how to identify and detect mirror websites effectively and quickly is very important.

2. Related work

The problem of text similarity and web deduplication should be focused on in the recognition and detection of website mirroring. The existing research is as follows.

2.1. Research of text similarity

Gerard Salton discussed the Vector Space Model (VSM) in 1969, which has been fully verified and applied [1]. The core idea of this model is to map the text into a multi-dimensional vector through a mapping algorithm, to perform the vector distance calculation on the space vector to realize the processing of complex information, and adopt the distance between the vectors to reflect the similarity between the documents. In 1988, Dumais et al. proposed a new information retrieval algebraic model - Latent Semantic Indexing (LSI) [2]. LSI is a special vector space model that uses statistics to obtain conceptual features for information retrieval, instead of traditional index characters and words. The LSI method does not involve natural language processing, and uses statistical methods to express the acquired features as specific concepts, which has high efficiency. Zhang et al. proposed the text similarity calculation method by the concept of Hamming distance. Huo et al. proposed a text similarity calculation method based on compressed sparse matrix vector multiplication, which has the advantages of simple calculation method and fast calculation speed of vector model [3]. In 1997, Antony et al., in response to the problems of some existing text similarity detection mechanisms at the time, proposed a copy detection mechanism, which can not only ensure that the text content of potential plagiarism can be fully detected, but also for the text between different topics. Krisztian et al. introduced MDR, a text overlap detection model, which can identify overlap or plagiarism in text[4]. Charikar proposed the Simhash algorithm in 2002 [5]. The Simhash algorithm is a representative application of locally sensitive hashing. The traditional hash function maps the text as evenly as possible across the entire space. A slight change in the text will cause the hash value to change. Partial-sensitive hashing can maintain the consistency of text and hash values, that is, the closeness of text content is reflected in the small distance between hash values, so that the distance of the hash value can indicate the degree of text
similarity. Simhash algorithm maps text to simple fingerprints, and obtains similar texts by comparing fingerprints, which greatly improves the retrieval speed and is suitable for massive text retrieval. Abdur et al. proposed a new text duplication detection algorithm based on ensemble statistics, called I-Mach [6]. Through the experimental comparison and verification of multiple data sets, I-Match can not only efficiently process long text content, but also will not ignore relatively short text information, and can well support plagiarism detection in large-scale text systems.

2.2. Research of page deduplication

Similar web page deduplication technology is derived from similar text detection technology, but it is not exactly the same as similar text detection technology. The similarity lies in the fact that most web pages belong to text, so the processes and methods of similar web page deduplication technology and similar text detection technology are very similar. The difference is: firstly, the webpages in similar webpage deduplication have no obvious paragraph structure, while the texts in similar text detection have clear paragraph structure; secondly, the webpages in similar webpage deduplication have webpage noise including headers, footers, and navigation columns and other parts that have nothing to do with the main information of the webpage.

The body content of the webpage is mainly obtained through certain algorithms after denoising the HTML tags of the webpage. Gomez et al. divide the similarity detection based on the content of web pages into two categories: cluster-based similarity detection and feature matching-based similarity detection [7]. Hofmann et al. proposed to conduct webpage similarity detection based on the vector space model (VSM, vector space model). The disadvantage of this type of algorithm is that the dimension of the vector model of the web page is too high, which makes the calculation time of the whole algorithm too long, and it is not suitable for a large number of web page similarity detection systems [8].

Based on link similarity detection is mainly to determine the similarity between web pages by comparing the in-links of the web pages (that is, the URL of the link to the web page) is consistent. For example, Shivakumar et al. proposed a hash-based URL detection method [9], Kapoor et al. proposed a Bloom filter-based URL detection algorithm [10], Jiang et al. proposed a deep learning-based URL detection algorithm, etc. [11]. This type of algorithm improves the performance of detecting malicious websites and reduces the storage of local duplicate data. However, its disadvantage is that when the webpage has fewer links, or when the webpage is a brand new webpage, similar webpages are difficult to detect.

In addition, typical similar web page deduplication methods include: Pedro Neves proposed to combine webpage deduplication and webpage prefetching technology to improve user-perceived delay [12]. Based on the principle of data cleaning, Lin proposed a feature code-based web page preprocessing to improve the effectiveness and efficiency of deduplication of similar web pages [13]. Wang et al. proposed a method to eliminate duplicate webpages based on feature codes by extracting information including the original website and webpage title [14]. Ricardo et al. proposed the DedupHTTP algorithm, which is based on a compressed cache method to achieve web page deduplication [15].

3. Theoretical research

This paper studies the webpage similarity detection algorithm based on Hamming distance and Simhash algorithm. The Simhash method are used to complete the preprocessing of the data, and then to process the high-dimensional data into a Simhash fingerprint (binary string) form that is easy to compare the similarity through operations such as extraction, weighting, merging, and dimensionality reduction.

3.1. Hamming distance

Hamming distance refers to the number of different characters at corresponding positions between two strings of equal length. The easiest way to calculate the Hamming distance is to perform an XOR operation on the two strings. For example, The Hamming distance between a strings \( x = (x_1, x_2, \ldots, x_n) \) and \( y = (y_1, y_2, \ldots, y_n) \) can be calculated according to the following formula.
Here, \( x_i \oplus y_i \) refers the exclusive OR operation between \( x_i \) and \( y_i \). Such as \( H(11001100, 10011001) = 11001100 \oplus 10011001 = 11010101 \).

The Hamming distance algorithm can simplify the calculation of long texts, is easy to understand and has higher calculation efficiency, is suitable for long text calculations, has greater calculation advantages, and can better improve the similarity calculation speed.

The similar webpage detection based on Hamming distance studied in this paper is also based on this theory. It converts high-dimensional data into lower-dimensional data that is easy to calculate similarity, thereby simplifying calculation complexity and saving query time. Different similarity calculation methods correspond to different data dimensionality reduction algorithms. The Simhash algorithm is used when the Hamming distance measures data similarity.

3.2. Simhash algorithm

The Simhash algorithm was first proposed by Moses Charikar in 2002. It was first applied in the deduplication system of Google web pages to solve the problems of content duplication and repetitive embedded advertisements in mirror sites [5][16]. Simhash algorithm has become one of the most effective algorithms in current text similarity detection processing. Similar to the general hash algorithm, the Simhash algorithm is essentially a dimensionality reduction technology. The high-dimensional data vector is mapped to a reduced-dimensional fingerprint, which can also be said to convert the natural language features of the text content into a binary digital form. Different from algorithms such as MD5 and SHA1, Simhash is a hash algorithm with local sensitivity. It can map the text content feature vector to a binary bit vector of a specified dimension, and the binary hash value represents the text content digital fingerprint.

We calculate the similarity of two Simhash values as follows:

\[
\text{sim}(S_1, S_2) = \frac{\sum_{k=1}^{128} (S_{1k} \oplus S_{2k})}{128}
\]

\( S_1, S_2 \) refers to two Simhash values, \( S_{1k}, S_{2k} \) refers to the values of the kth bit of each string, 128 is the number of bits in the binary string obtained after the hash function used in this article.

The calculation process of the Simhash algorithm is as follows:

- **Initialization**: Create an n-dimensional vector, and initialize all dimensions of the vector to 0.
- **Stem operation**: perform word segmentation processing on the input file, use Ported stemming analyzer to extract stems and convert all words into lowercase form.
- **Weight distribution**: using the weight calculation method, each feature word is assigned a weight.
- **Hash processing**: Use traditional hash functions such as MD5 and SHA-1 for hash processing. This type of algorithm can generate a unique digital fingerprint and is very sensitive to the input data. If the file is slightly modified, its MD5 value will change dramatically.
- **Weighting**: Through the above steps to generate the hash result, it is necessary to form a weighted number string according to the weight of the word. If the hash value is 1, the weight value of the word is increased, and if the hash value is 0, the weight value of the word is subtracted, as shown in equation 3.

\[
f(\text{hash}_w) = \begin{cases} 
\text{weight}_i & \text{hash}_w = 1 \\
-\text{weight}_i & \text{hash}_w = 0 
\end{cases}
\]

- **Combine**: Accumulate the sequence values calculated of each feature word above and merge them into a sequence string, recorded as \( T_j \), see equation 4.
Dimensionality reduction: Turn the sequence string calculated in step (6) into a binary string to form the final Simhash value. Bits greater than 0 are recorded as 1, and bits less than or equal to 0 are recorded as 0, see equation 5.

$$T_i = \begin{cases} 1 & T_i > 0 \\ 0 & T_i \leq 0 \end{cases}$$

Simhash is a local sensitive hash algorithm (LSH). Simhash generates fingerprints based on the content of the file. Due to the nature of LSH, texts with similar content will produce similar Simhash fingerprint values.

4. Experiment and analysis

This article carried out a series of experiments, using the Simhash algorithm to generate a fingerprint for each web page, and using the Hamming distance to calculate the similarity. Here, the authors select three webpage detection methods to detect 6 test mirror websites, and compare the performance of the three algorithms. The three algorithms are segmented signature algorithm, BL method and Simhash method based on Hamming distance proposed in this paper.

The performance evaluation standard of the detection algorithm is generally described by the PR curve, where P refers to the accuracy rate (see equations 6), in this experiment, it refers to the proportion of the results that are judged correctly; R refers to the recall rate (see equation 7), which measures the correctness of the detection. The result covers the proportion of all positive samples, so it is also called recall. The parameter definitions defined in the formula are shown in Table 1.

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

|                | True(T) | False(F) |
|----------------|---------|----------|
| Positive(P)    | TP      | FP       |
| Negative(N)    | TN      | FN       |

It can be seen from the figure 1 that compared to the other two methods, the method proposed in this paper has higher precisions and recall rates. In addition, it can be concluded from the experiments: the text is mapped to a locally sensitive hash fingerprint, so that similar text fingerprints are the same or similar, reflecting the characteristics of the text; the algorithm actually performs feature dimensionality reduction on the text. The long text is represented by a 128-bit binary string. For large-scale texts, the fingerprints occupy a small space after feature dimensionality reduction, so that the same feature text can be quickly searched; by establishing a segmented inverted index method, similar fingerprints can be retrieved quickly, and the range of expression of similarity can be expanded; calculating the Hamming distance of Simhash can indicate the degree of similarity between texts, and the algorithm has higher efficiency.
5. Conclusion
The improved webpage de-duplication algorithm based on Simhash proposed in this paper converts the
text collection into Simhash fingerprints for storage, and calculates the similarity of the two fingerprints
by calculating the Hamming distance, thereby obtaining the similarity of the webpage. Experiments
show that the algorithm has higher precisions and recall rates, and can be better applied to the
identification and detection of malicious mirror websites. Due to the various forms of similar webpages,
especially some malicious mirror websites will make corrections in the source code, format, etc., and
even set up anti-detection programs, which brings great difficulties to the identification and detection of
similar webpages. Future research work will focus on the improvement of the algorithm.

Acknowledgments
In this paper, the research was sponsored by the Basic Public Welfare Research Program of Natural
Science Foundation of Zhejiang Province under Grant No. LGF19F020006 and NSFC program under
Grant No.62002362.

References
[1] Salton, G., Wong, A., Yang, C.S. (1975) A Vector Space Model for Automatic Indexing.
Communications of the ACM, 18:613–620.
[2] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R. (1990) Indexing by
latent semantic analysis. Journal of the Association for Information Science & Technology,
41 (6): 391-407.
[3] Huo, H., Feng, B. (2005) Document similarity degree measuring based on compressed sparse
matrix vector multiplication technique. Mini-micro Systems, 26 (6): 988-990.
[4] Monostori, K., Zaslavsky, A., Schmidt, H. (2000) Documer overlap detection system for
distributed digital libraries. In: Proceedings of the fifth ACM conference on Digital libraries.
San Antonio. pp. 226-227.
[5] Charikar, M. (2002) Similarity estimation techniques from rounding algorithm. In: 34th Annual
Symposium on Theory of Computing. Rome. pp. 380-388.
[6] KOŁCZ, A., Chowdhury, A. (2008) Lexicon randomization for near-duplicate detection with
I-match. Journal of Supercom-puting, 45(3): 255-276.
[7] Gomez, N. (2014) Similarity preserving snippet-based visualization of Web search results. IEEE
Transactions on Visualization and Computer Graphics, 20(3): 457-470.
[8] Hofmann, T. (2017) Probabilistic latent semantic indexing. In: ACM SIGIR Forum. Tokyo. pp.
[9] Shivakumar, N., Garcia, M.H. (1998) Finding near-replicas of documents on the Web. In: International Workshop on the World Wide Web and Databases. Valencia. pp. 202-212.

[10] Kapoor, A., Arora, V. (2016) Application of bloom filter for duplicate URL detection in a web crawler. In: IEEE International Conference on Collaboration and Internet Computing. Pittsburgh. pp. 246-255.

[11] Jiang, J., Chen, J., Chook, K.R. (2017) A deep learning based online malicious URL and DNS detection scheme. In: International Conference on Security and Privacy in Communication Systems. Niagara Falls. pp. 438-448.

[12] Perdo, N., Paulo, F., João, B. (2013) Leveraging Web Prefetching Systems with Data Deduplication. In: IEEE 12th International Symposium on Network Computing and Applications. Cambridge. pp. 259-262.

[13] Lin, J.M., Liu, D.S., Gao, S.W., Chen, W. (2009) A Web Page De-duplication Algorithm Based on Data Clearing. In: International Joint Conference on Artificial Intelligence. Hainan Island. pp. 544-547.

[14] Wang, M.Y., Liu, D.S. (2009) The Research of Web Page De-duplication Based on Web Pages Reshipment Statement. In: The First International Workshop on Database Technology and Applications. Wuhan. pp. 271-274.

[15] Ricardo, F., Joao, B. (2011) End-to-End data deduplication for the mobile web. In: IEEE 10th International Symposium on Network Computing and Applications. Cambridge. pp. 334-337.

[16] Sam, B., Shaabi, M., Sam, O., Luke, S., Mark, S., Judita, P. (2012) Two approaches to Semantic text similarity. In: First Joint Conference on Lexical and Computational Semantics. Montreal. pp. 655-661.