Research on Algorithm for Informational Text De-Duplication

Linqing Deng[1], Yingying Li[1,2], Jie Hu[1]
[1] School of Software, Shanxi Agricultural University, Jinzhong 030801, China
[2] Mixed and Virtual Reality Research Lab, Vicubelab, Faculty of Engineering, Universiti Teknologi Malaysia (UTM), Johor Bahru, Johor, Malaysia

Abstract: In recent years, with the continuous development of China's science and technology and computer science and technology, the technology of transmitting various information in the form of such text as Chinese and short text has been developing and popularizing, like Micro-blog, and WeChat official account. The continuous increase in the dissemination of short text information provide various resources for information decision-making and information, but there has also been a large amount of redundancy, especially in the case of invalid and repetitive information for informational texts. In such a large and repetitive information set, the storage capacity of the system is heavily occupied, which is not conducive to the collection and extraction of effective information and data from informational texts, seriously affecting the accuracy of information decision-making and affecting the timeliness of information. Therefore, it is necessary to strengthen the research on method for informational text de-duplication in this context. Taking informational text de-duplication as an example, this paper analyzes the current research on technology for text de-duplication at home and abroad, and conducts research on methods for informational text de-duplication based on relevant technologies, in order to provide certain reference ideas for enterprises when carrying out informational text de-duplication.

Keywords: Informational Text; De-Duplication Algorithm; De-Duplication Technology

1. Introduction

In the new era, with the arrival of the big data era, digital information is showing a rapid development trend. With the assistance of information technologies such as cloud computing and big data, the amount of data is experiencing explosive growth. As more and more data is generated and stored globally, the demand for data storage capacity is increasing. However, regardless of the way data is stored, there will be a large amount of redundancy, resulting in a large amount of duplicate data in many systems and occupation of storage capacity. At present, more and more research is paying attention to data redundancy, and reducing data redundancy can effectively improve data storage capacity. Especially for text data files, deleting duplicative text and data can effectively improve text storage capacity, so de-duplication technology has emerged. The core of strengthening the informational text de-duplication method is to store the unique data object of the article and optimize other duplicative articles and data. At present, the most common de-duplication algorithm technologies mainly include k-shingle, Min Hash, and Sim Hash algorithms. By adopting these technologies, the problem of text duplication can be effectively improved, achieving effective text duplication [1].

2. Research on Text De-Duplication Technology at Home and Abroad

In the research on text de-duplication technology both domestically and internationally, it is divided into two categories based on its algorithm principles. One method is based on text syntax and adopts string comparison methods to search for similar text in large-scale file systems. Although this method does not involve text de-duplication related technologies, it proposes an idea of de-duplication detection, providing ideas for developing text replication detection systems, and a basic detection framework has been

established. In addition, grammar-based duplicative text detection can directly search for relevant strings in the document through string matching algorithms, matching and detecting duplicative text. The other method is mainly based on semantics, which establishes a detection system based on vector space models through statistical word frequency, and compares and analyzes relevant sentences through detection. In the detection process, based on semantics, documents are decomposed into tree structures, and vector dot product method is used to verify document similarity, effectively improving the speed and efficiency of document repeated verification. It can be seen that there is a deep understanding of text de-duplication technology both domestically and internationally. For similar Chinese informational texts, the main reference for developing de-duplication technology is the English text de-duplication method. Applying relevant de-duplication technology to Chinese web page duplication checking can have a certain effect. However, with the increasing content and information of informational text, many Chinese informational text information has not been effectively utilized, resulting in omissions when using relevant de-duplication techniques. It proves that there are still many imperfections in the current research on informational text de-duplication technology. Due to the massive characteristics and concise content of informational texts, segmented de-duplication algorithms can be used when using de-duplication techniques [2].

Based on this situation, research on informational text de-duplication methods can use Bloom Filter, Tire Tree algorithm, and Sim Hash algorithm, which can effectively improve text detection efficiency and enhance text de-duplication accuracy.

3. Research Model of Informational Text Re-Duplication Method

When conducting de-duplication detection on Chinese text, it is necessary to fully recognize the characteristics of Chinese text, especially for informational text, it is brief and massive. To effectively distinguish the differences between Chinese and English, when studying the model of informational text de-duplication methods, informational text features can be the main focus, and the idea of feature code extraction algorithms can be used to design Chinese informational text de-duplication algorithms, and combine the retrieval time and complexity to study its memory allocation, in order to improve the accuracy of de-duplication. During this process, it can be optimized based on Bloom Filter and Tire tree algorithms. However, the use of Bloom Filter and Tire tree algorithms can only remove completely duplicate parts of informational short text, and cannot effectively filter and remove similar parts. It results in a small amount of text with similar content still remaining after reprocessing in informational text, which affects the originality of the text to some extent. For example, when calculating informational text, after making simple modifications to the content of others' information, similar text cannot be detected during original statistics, resulting in the content being treated as original informational text, as well as causing unnecessary misunderstandings. In order to perform similar text de-duplication on informational text after complete de-duplication, the Sim Hash algorithm can be introduced. It can be seen that in the study of informational text de-duplication methods, the main models are extracting relevant datasets of informational text and pre-processing them. The datasets are first processed through Bloom Filter or Tire tree to completely remove duplicates in the text. Then, the Sim Hash algorithm is used to perform similarity and duplication removal on informational texts. Finally, the final de-duplication article will be obtained [3].

4. Design of Basic Algorithm for Informational Text De-Duplication

Author names and affiliations are to be Based on the practical application of informational text, the relevant content mainly presents semi structured and unstructured features. Therefore, there are a large number of completely repetitive texts and similar problems in the relevant data of informational texts. For similar texts, it takes more time and space to compare the texts, so it is generally not possible to use a large-scale direct comparison method when processing similar texts. For this type of text, in order to effectively remove duplicate and similar informational texts,

Bloom Filter and Tire tree algorithms can be used to remove a large number of completely duplicative texts, and then Sim Hash algorithm can be used to remove similar repetition from informational texts, improving the efficiency of informational text de-duplication [4].

4.1 Design of Bloom Filter Algorithm

Bloom Filter is an algorithm based on a hash function that quickly detects whether there are duplicative texts in the datasets by searching for relevant data in the function. The hash function involved in Bloom Filter can effectively save space and improve storage efficiency compared to ordinary hash tables. However, the Bloom Filter algorithm has a certain degree of fault tolerance and may have some false positives when retrieving texts, such as data exists (possibly with false positives) and data is not in the set (necessarily not). Therefore, when de-duplicating informational text, the use of Bloom Filter algorithm can only simply remove a large amount of completely duplicative text. So it is necessary to cooperate with other algorithms to improve the accuracy of text de-duplication [5].

4.2 Design of Tire Tree Algorithm

The Tire tree refers to the dictionary tree. The use of dictionary trees can achieve higher query efficiency. Therefore, when removing duplicative informational texts, dictionary trees can be used to filter and match duplicative texts through string search. The idea of deduplication with Trie Tree algorithm is to exchange space and use the common prefix of strings to find duplicative text. When using Trie tree algorithm, the root node used does not contain characters. From the root node to a certain node, the characters passing through the path will be meaningfully retrieved. Due to the different sub node characters included in each node, duplicative text can be more intuitively detected during the de-duplication process.

4.3 Design of Sim Hash algorithm

Sim Hash is actually a dimensionality reduction technique that uses relevant programs to input a vector and obtain a value through operations. In the process of processing informational text data, the input vector is a set of feature vectors of informational text. When selecting feature values, the more effective the selection, the higher the accuracy of text de-duplication. Due to the brevity of informational texts, several consecutive word strings can be selected when selecting feature values and inputted into Sim Hash for operation. By comparing the correlation values of two text signatures, the similarity between the two texts is analyzed, and a pre-set similarity threshold is used to analyze the relevant texts, and similar texts in informational texts are filtered and effectively removed [6].

5. Conclusion

In summary, this paper conducts research on text redundancy de-duplication technology at home and abroad. Combining several commonly used text de-duplication technologies and taking informational text as an example, it studies text duplication detection and de-duplication algorithms, analyzes the characteristics of Bloom Filter, Tire Tree algorithm, and Sim Hash algorithm. Among them, Sim Hash algorithm, compared to other algorithms, has the characteristics of fast processing speed and high accuracy of results. When detecting informational text, it can be effectively applied to similar text detection and redundant data de-duplication. By continuously optimizing the Sim Hash algorithm, it can better improve its application efficiency in text de-duplication and further improve the efficiency of related technologies in detecting duplicate text.

Acknowledgment

We would like to express our deepest gratitude and appreciation to the project "Construction of a Large-Scale Scientific Research Information Sentence Database and Massive Text Reduplication Research" (project number: 2020QC07), which was initiated by Shanxi Agricultural University in 2020. This project has provided important resources and financial support for our research and has had a positive impact on our research work.

References

[1] Wang Jinyun, Xiang Yang. Research on Text Semantic De-duplication Algorithm

Based on Keyword Graph Representation [J/OL]. Computer Applications: 1-8.
[2] Zhang Yanan, Chen Weiwei, Fu Yinjin, et al. Research on Text De-duplication Algorithm Based on Sim Hash [J]. Computer Technology and Development, 2022, 32 (08): 26-32.
[3] Yao Qingfeng. Research on Text Processing and Mining Algorithms for Social Media [D]. Beijing University of Posts and Telecommunications, 2022.
[4] Wang Tiannan, Feng Feng. Research on Text Similarity Detection Algorithm Based on Sim Hash [J]. Electronic Testing, 2019, (15): 87-89.
[5] Zhang Hang, Sheng Zhiwei, Zhang Shibin et al. The Application of Sim Hash Algorithm in Text De-duplication [J]. Computer Engineering and Applications, 2020, 56 (11): 246-251.
[6] Cai Yanjing. Research on Text Similarity Deparallelization Algorithm [J]. Electronic Production, 2018(10): 35-37.

