An Improvement Method of Duplicate Webpage Detection

Chengqi Zhang[1, a], Wenqian Shang [2,b] and Yafeng Li [3,c]

[1]Department of Computer Sciences, Communication University of China, China

[2]Department of Computer Sciences, Communication University of China, China

[3]Department of Computer Sciences, Communication University of China, China

[a]email:18201017652@163.com,[b]email:shangwenqian@cuc.edu.cn

Keywords: search engine; duplicate detection; BloomFilter; Fuzzy Hamming distance

Abstract. As Internet is very easy to implement the diffusing and sharing of resources, duplication of pages on the Internet is very large. The search engine as an index tool of Internet resources is facing a serious repeat testing, its crawler will encounter a large number of links of duplicate content. If these links are all added to the download queue, it will cause a serious drop in performance and this would seriously affect the user experience. In this paper, we adopt an improved duplicate detection method------using BloomFilter combining with fuzzy Hamming distance. This will not only meet the detection of duplicate content, but also it will meet the needs of users.

Introduction

With the massive popularity of the Internet, a variety of resources show a geometric explosion of growth. The ever-growing Internet promotes the continuous development of search engine technology and makes it become the primary means of accessing information from the Internet. However, because of the information reproduced very easily on the network, it results in a large number of duplicate pages filled in the Internet. Search engine at work is facing a serious problem of duplicate content. Its spiders will get a large number of duplicate links. If these links are all added to the download queue, it will greatly increase the pressure of the search engine to download these links, and if these duplicate pages are added to the index module in the user query, it will return to the user a lot of duplicate pages. These pages may be the fully repeated verbatim, may be part of the repeat. This will increase the burden of the search engine and the burden of users to browse. This will waste a lot of storage resources and reduce the index of efficiency. Therefore, accurate and fast to remove duplicate pages is undoubtedly one of the key technologies to improve the quality of the search engine.

The existence of duplicate information has the following drawbacks: seriously hurt the user experience; easily lead to the dead links and spam; waste storage space and difficult to update, etc. The current duplicate detection method is very limit. Search engines that users most commonly use always give duplicate results. Solve the problem of duplicate pages can be more effective to provide the required resources. Facing vast amounts of information, if it only relies on the manual processing, it will be human resources consuming and time consuming. It cannot meet the actual needs. How to automatically detect duplicate content in a data set becomes a hotspot of researchers in recent years. In this paper, we give a new method of web page duplicate detection. We combine Bloom filter and fuzzy Hamming distance, greatly reduce the number of repeats, and improve the user experience.

Status of the duplicate detection method

Homologous web page detection

This situation is lead by the interconnection between the pages. In our experiments, the ratio of duplicate links and the whole links can reach 6:7. If adding these duplicate links to the pending download queue, it will give the download module several times pressure. Because of the huge

© the authors

number of links, we must find an effective way to accommodate such a huge number of links to the duplicate detection.

Completely duplicate pages

Between the duplicate pages of this type is a verbatim copy. Including FAQ (Frequently Asked Questions), RFC (Request For Comments), legal documents, hot news, and user-reprinted artic.

Part of the duplicate pages

The duplicate pages of this type are not a complete copy of the original page, but will alter some of the content and then re-emergence on the Internet. Generally not the same in terms of site templates, web page format, the site administrator signature, the subject matter of the body part of the page can be partially changed. For example, the text of the updated news and articles reproduced in part, the existing social networking sites, a large number of partially modified articles make the types of duplication become more serious.

With the steady increase of the pages on the Internet, the duplicate pages become more and more. Duplicate pages have become one of the important issues that affect the accuracy of search engine results. Quickly and accurately find similar pages will improve search engine's quality of service, and improve the accuracy of the search results.

Here are the commonly used algorithms of duplicate detection:

A. Different methods for dealing with replicas of Web Pages

The method based on the database. All the links are stored in the database. Getting rid of duplicate url needs to traverse the database. This method is stability and security, but the efficiency and speed is very low. This method applies to the system that requires high security.

The method based on Memory. The all links are written to the memory hash table, and then through Hash function it can find the corresponding number of bits to achieve the purpose of duplicate detection. Although this method is better than the repeated testing of the database on the efficiency and speed, but because of the limited memory space, this will lead to be paralyzed. Its structure is simple and fast, so this way is very suitable for small focused crawler.

The method based on Bloom Filter. Bloom filter is proposed in 1970 by Bloom Barton. Its main idea is using a 16 times larger address space, so that all eight hash function mapped to the address space inside. Its advantage is the fast speed and space-saving, but it appears the case of judgment error, the common approach is to create a white list, storage the possible miscarriage of justice.

B. Duplicate detection methods according to the page content

Syntax-based method. It adopts the string comparison methods. This method requires segmentation of the document, selecting some strings, these strings are called "fingerprint". Fingerprint is mapped into the hash table. A fingerprint corresponds to a number of the same fingerprint in the final statistics hash table number or ratio, as the text similarity basis. For example, DSC and DSC-SS algorithm proposed by Broder [1] and the I-Match algorithm [2].

Semantic-based approach. This is the word frequency statistics. The number of occurrences of each word in each document of such method should be statistical, based on word frequency to constitute the document feature vectors, using dot product, cosine or similar means to measure the feature vectors of two documents. Such as the algorithm bases on vector space model (VSM) that is first proposed by Gerard [3].

Form the above situation, we integrate the advantages of the various methods and to propose a comprehensive method that using the Bloom filter can improve for homologous web page. For exactly the same and part of the same page, we take fuzzy Hamming distance method. The experimental results show that it is effective.

Key technology

Bloom Filter

Bloom Filter is an excellent data structure for succinctly representing a set in order to support membership queries [4].

The basic idea of this algorithm is:

© the authors

Set the data set A as $\{a_1, a_2 ... a_n\}$, containing n elements.

Bloom Filter using a length of m-bit vector V to represent the elements in the collection, the bit vector is initialized to 0.

k hash functions $h_1, h_2 ... h_k$ that has uniform distribution.

For the adding operation of the elements, first using k hash function generates k random numbers $h_1, h_2, ..., h_k$, then sets the corresponding vector bit $h_1, h_2, ..., h_k$ of V to 1; Similarly, searching of the elements is to determine the appropriate bits are all 1. During the collection of elements of the Bloom Filter algorithm expressed by the k hash functions is the bit vector corresponding position to 1. After many collection element to increase the operation, the bit vector to be repeated is set to 1.

Error rate estimates: There will be a certain size of the error rate (false positive rate) when we use Bloom Filter, we estimated the error rate of Bloom Filter to judge whether an element belongs to a collection it represents. To simplify the model estimated before, we assume that $k_n < m$ and each hash function is completely random. When the set $S = \{x_1, x_2, ..., x_n\}$ of all elements of k hash functions is mapped to the median of the m-bit group, this median group or a 0 probability is:

$$p' = \left(1 - \frac{1}{m}\right)^{kn} \approx e^{-kn/m} \tag{1}$$

We let $p = e^{-kn/m}$, and p is a convenient and very close (within O(1/m)) approximation for p'.

Now, let ρ be the proportion of 0 bits after all the n elements are inserted in the table. The expected value for ρ is of course E (ρ) = p', Conditioned on ρ, the probability of a false positive is

$$(1-p)^k \approx (1-p')^k \approx (1-p)^k \tag{2}$$

We already discussed the second approximation. The first one is justified by the fact that high probability ρ is very close to its mean. We will return to this fact at the end of this section.

We let:

$$f' = \left(1 - \left(1 - \frac{1}{m}\right)^{kn}\right)^k = (1-p')^k \tag{3}$$

And

$$f = \left(1 - e^{-kn/m}\right)^k = (1-p)^k \tag{4}$$

In general, it is often easier to use the asymptotic approximations p and f in analysis, rather than p' and f' [5][6].

B. Fuzzy Hamming distance

The approach is designed as follows:

Downloading pages; extracting text; doing text preprocessing; deleting stop words; extracting of feature vectors; sorting according to word frequency.

Suppose a text feature vector is C $(x_1, x_2 ... x_n)$, to be detected text feature vector is D_i $(y_1, y_2 ... y_m)$. Through calculating the fuzzy Hamming distance about it, we can find the radius of the neighborhood of the collection D. If it finds with the same eigenvalue, the eigenvalue corresponding bit can be replaced, then K does not count if they can not find the count K + when N <M, the finally K plus (M-N)/N, if N> M, it will be the last to get the K plus (N-M) / N. That is:

$$\beta = \frac{k}{N} \tag{5}$$

If the value is greater than a certain threshold, the text is not the same or as similar text.

Experimental results and analysis

The article in the experiment is varying degrees of change, record the changes in K / N values, as shown Fig. 2:

Published by Atlantis Press, Paris, France.
© the authors

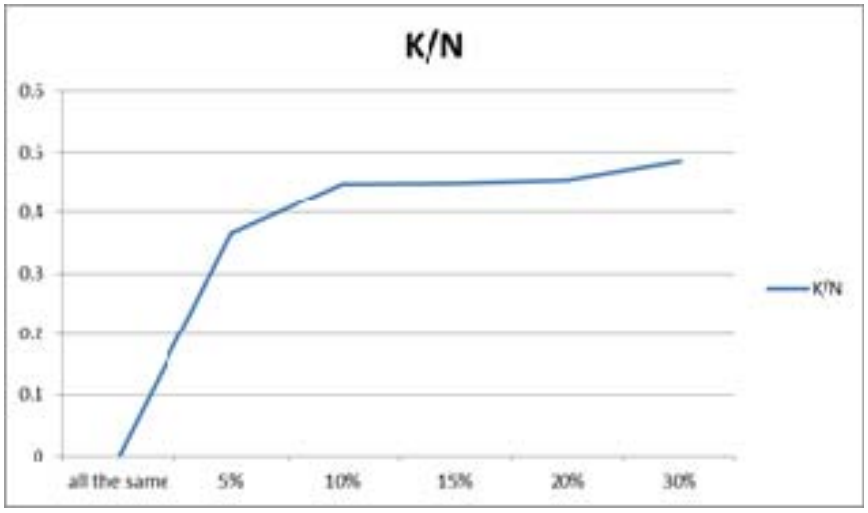

Fig.2 The value of K/N

The search engines can define the threshold as Fig. 2 according to their needs, which can meet the diverse needs of search engines.

Conclusion

The voluminous amounts of web documents swarming the web have posed a huge challenge to the web search engines. This makes them render results of less relevance to the users. In this paper, we propose an improved duplicate detection method. This method adopts fuzzy Hamming distance combining with Bloom Filter. The experimental results show better performance. It can meet the diverse needs of the work. In the future, we will further improve this method.

Acknowledgment

This paper is supported by the 48th postdoctoral foundation of China (20100480357).

References

[1] N. Shivakumar and G. Olian. Building a Scalable and Accurate Copy Detection Mechanism. 1st ACM Int.Conference, p. 160-168 (1996).

[2] A. Chowdhury and F. O. Rieder. Collection Statistics for Fast Duplicate Document Detection. ACM Trans. Information System, 20(2), p. 171-191.

[3] G. Salton and A. Wong. A vector space model for information retrieval. Communications of the ACM. Vol. 18(1975), p. 613-620.

[4] B. Bloom. Space/time tradeoffs in hash coding with allowable errors. Commun. Of ACM, vol. 13(7)(1970), p. 422–426.

[5] A. Broder and M. Mitzenmacher. Network applications of bloom filters: A survey. Internet Mathematics, p. 485–509 (2005).

[6] V. A. Narayana, P. Premchand and A.Govardhan. A Novel and Efficient Approach For Near Duplicate Page Detection in Web Crawling. IEEE Transl. J. Magn. India. p.1945-1945 (2009).

© the authors