The Chinese Duplicate Web Pages Detection Algorithm based on Edit Distance

Junxiu An
Chengdu University of Information Technology, Chengdu, P.R.China
Email: anjunxiu@cuit.edu.cn

Pengsen Cheng
Chengdu University of Information Technology, Chengdu, P.R.China
Email: cps11@163.com

***Abstract*—On one hand, redundant pages could increase searching burden of the search engine. On the other hand, they would lower the user's experience. So it is necessary to deal with the pages. To achieve near-replicas detection, most of the algorithms depend on web page content extraction currently. But the cost of content extraction is large and it is difficult. What's more, it becomes much harder to extract web content properly. This paper addresses these issues through the following ways: it gets the definition of the largest number of common character by taking antisense concept of edit distance; it suggests that the feature string of web page built by a previous Chinese character of period in simple processing text; and it utilizes the largest number of common character to calculate the overlap factor between the feature strings of web page. As a consequence, this paper hopes to achieve near-replicas detection in high noise environment, avoiding extracting the content of web page. The algorithm is proven efficient in our experiment testing: the recall rate of web pages reaches 96.7%, and the precision rate reaches 97.8%.**

***Index Terms*—Near-replicas detection, edit distance, the largest number of common character, feature string of web page**

I. Introduction

The duplicate web pages and near-duplicate web pages are increasing on the Internet, because of the Internet users' coping, altering the web pages at discretion, secondary integrating and modifying the articles. For the large duplication of the information, it is difficult to improve the precision of search results and it is also impossible to provide the user high experience services. And it would increase the burden on the search engines. For example, the web pages which contain the same information with different links would reduce the query efficiency of inverted index, and increase the maintenance costs of inverted index. Therefore, to identify duplicate and near-duplicate web pages, it provides necessary basis in the process of the Internet ordering to achieve information removing and integration, and it increases the recall rate and precision rate of search engines by raising effective information content in search result.

All this, rapid development of the Internet technology, fierce competition among the Internet companies and increasing requirements for the Internet service users, contribute to the web page noise rising rapidly that the layout of a web page from simple to heavy and complicated. The costs of page content extracting are out of proportion to the precision of extracting. That is the reason why classical algorithm is failure and the complexity of the current algorithm is increased. This paper selects the feature string coming from the body of the web page content ranges as far as possible by analyzing the component of web page. To avoid the extraction of the body of the page content, it introduces edit distance algorithm to decrease the impact of noise.

II. Related Work

Boder put forward the DSC((Digital Syntactic Clustering) to detect the near-duplicate web pages[1]. In the DSC, a paper is composed by several shingles, and it decides the duplicate pages by calculating the number of same shingles in text. This algorithm has fewer comparisons than comparison of full-text, but its efficiency is low. Later, Boder put forward an improved algorithm named DSC-SS(Super Shingle)[2]. The DSC-SS merges several shingles to a super shingle, and it calculates the hash value of the super shingles. But the efficiency is still lower to deal with the large scale web pages.

Narayanan Shivakumar and Hector Garcia-Molina offer a block signature based approach which then is use for Google[3]. The algorithm partitions a paper to chunks in unit of word, sequence, sentence, paragraph or full text. A paper is expressed by several 32-bit hash values which are output of every single chunk. The advantages of the algorithm are flexible partition and efficiency querying. And the disadvantage is it needs to update index frequently.

The I-Match filters the shingles [4], and it gets a MD5 value by inputting shingles. The signature results are unstable, for the shingle frequency in all documents is unstable. Based on I-Match, A. Kołcz and A. Chowdhury offer an improved I-Match with lexicon randomization to raise the stability of signature [5].

doi:10.4304/jsw.8.7.1666-1670

Wu PingBo put forward a duplicated web pages of Chinese algorithm based on string of feature code[6]. The way not only utilizes page content but also the structure of the page text. But the difference order of the text sections or loss of paragraphs has a great influence on the result.

Wei LiXia and Zheng JiaHeng study a text structure based approach[7]. The main idea is to generate a structure tree relying on the page structure. The way has high recall and precision on the mirroring pages detection. But it is complex and lower efficiency. The most important is it needs to maintenance the structure trees in the larger space.

A text structure and extraction of long sentences based approach [8] is suggested by Hang Ren at al. The way extracts the feature from the content dynamically and it calculates the fingerprint hierarchically to ensure the efficiency. It gets the node fingerprint to ensure the robustness based on the algorithm of long sentences extraction. But the way is complex and expensive on memory.

Cao Yujuan put forward a concept and semantic network based approach[9]. This algorithm has a better time and a space complexity, and it does not rely on the corpus. But the key concept identification is difficult in the short page processing. It reduces the recall and precision.

Cheng and An forward a feature words based approach[10] to detect the duplicate news web pages. The algorithm constitutes the feature words group by picking up the highest frequency words from seven categories of part-of-speech. The disadvantage of the algorithm is it only can detect the article which is written in standard grammar.

The literature [11] tests the effectiveness of three early duplicated detection algorithms -- approximate string matching, space vector and SCAM, and it applies these algorithms to the OCR duplicated detection in the Internet. The literature [12] analysis the detection way and technology of various text copy detection systems before 2003, and it compares the similarities and differences in the key technology of this systems. The literature [13] studies the interaction between several important parameters of the duplicate document recognition algorithm. It divides the detection ways into shingle based approach and term based approach, and it points out that the selection criteria and basis of common hash functions used in algorithm. The literature [14] analysis and summarizes the present research situation of the duplicated web pages detection technology in China.

The detection algorithm based on different principle can be divided into three classes. The first class is that it partitions a paper to several special units. Then these algorithms compare the units or the hash values of the units. This class can be easily implemented, but its efficiency is low and it is easily affected by noise. The second class is that it constructs a structure tree of every page and compares the structure trees. These algorithms have higher recall and precision, but they are complicate to implement and the efficiency is low. The third class is that it extracts the feature of the page and compares the features. These algorithms can be easily implemented and its efficiency is high, but how to accurately extract the feature is hard. If the algorithm were not perfect, there would be noise in the feature.

III. The Definition of Duplicated and Near-duplicated Pages

It is assumed that there are two articles A and B which represents the paragraph collection respectively. The expression $C = A \bigcap B$ means the intersection of A and B, which exclude individual words or symbols differences arising differences between the paragraphs. When the pages which A and B represents are duplicate or near-duplicate pages, the possible values of C shown in figure 1:

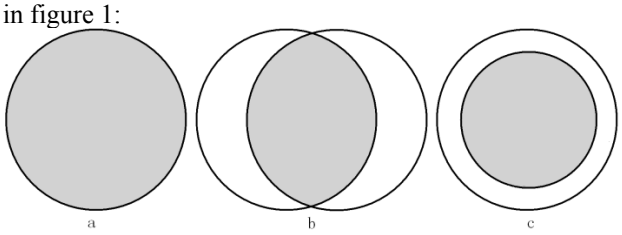

Fig.1. The possible result of intersection of duplicate or near-duplicate web pages

The shaded area in figure 1 represent the possible values of C. The class *a* indicates the paragraph content of the page which represented by A and B are all same. When the content is shorter, the phenomenon is most common. The class *b* indicates the part content of A and B is same. In the actual web pages, this situation is relatively complex. The main reasons are: 1.different interpretation of the same event, for example, the beginning of the article applied the specific content of the legal regulations which is new introduced, and the end of article analyze and discuss it; 2.when reprint the article, people modify it; 3.just caused by reprinted noise; 4. same main content but different reply in the forums or blogs. The class *c* indicates the B is part of A. The common scenario is the different page way or just picked up the part of the original page.

In three classes *a*, *b* and *c*, it is easy to identify the *a* and *c*, but the class *b* is the articles which are different expression under the same topic. If the granularity of recognition is larger, it can classify all articles as a class. If the granularity of recognition is smaller, it is difficult to distinguish all subclass under the *b*.

IV. The Design of Edit Distance Based Approach

The content of article is important in page detection. The feature strings which are extracted from an article must reflect the content in the possible shortest length. Because of the page layout or human modifications, there would be some kinds of noise in the duplicate pages. The consistency of feature string extraction rules will lead to noise contained in the feature string inevitably. In this case, it needs some kind of evaluation rules to reduce the sensitivity of algorithm to the noise.

A. The Minimum Edit Distance

The edit distance[15] is used to compare the similarity of two strings, and it was raised by levenshtein in 1966. The basic idea is that smaller the distance between two strings, the more similar the two strings.

It is assumed that the n is the length of string $Sour$, and the m $(m \geq n)$ is the length of string $Dest$.

Definition 1: The minimum edit distance $edit(Sour, Dest)$ means the minimum number of conversion operations that change string $Sour$ to string $Dest$. There are three conversion operations:

Change a character;

Insert a character;

Delete a character.

There is a following properties between n, m, and $edit(Sour, Dest)$:

Property 1: $m - n \leq edit(Sour, Dest) \leq m$.

The definition one explains that the minimum number of operations needed when change the shorter string to the longer string. The property one indicates the range of minimum edit distance from two special cases: the shorter string is a subsequence of the longer string, and the shorter string and the longer string are totally different. The minimum edit distance is calculation the changing number of characters, and the unchanging number of characters can come through the minimum edit distance calculation. Therefore, it gets following definition:

Definition 2: The largest common number of characters $LCC(Sour, Dest)$ means the maximum number of characters no need to change when change $Sour$ to $Dest$ with the minimum edit distance, $LCC(Sour, Dest) = m - edit(Sour, Dest)$

It gets the following property from the property one and the definition two:

Property 2: $0 \leq LCC(Sour, Dest) \leq n$

B. The Extraction of Feature String

Generally, there are following sections to compose a public Chinese web page: navigation bar, text, comments, related links, advertisements and copyright information. Through analyzing the Chinese web pages, it founds that the navigation bar is made up of several individual word; the related links are simple words or article titles; the advertisement also is a simple sentence; the copyright information is made up of two to three short sentences. The period which is at the end of a sentence is very rare, or does not appear in these parts, and the period is a high frequency of punctuation in Chinese. Therefore, the period is a symbol of character strings.

As shown in figure 2, in the character string extraction, it only extracts a Chinese character before the period. If there is not a Chinese character before the period, such as other punctuations, English character or digit, it will not be extracted.

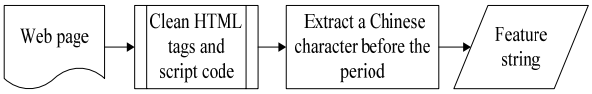

Fig.2. The process of extraction a web page feature string

The HTML source code is written in a certain order. The content from different part is not be nested. So, after cleaning the web label, the saved text from different part cannot be nested. It reflected in feature string is the Chinese words coming from same part are together.

C. The Design of Evaluation Function of Multiplicity

The extraction of feature strings is based on containing content and out of content to implement. There may be some noise in feature strings. Therefore the study uses the largest common number of characters to measure the fuzziness of web page.

Definition 3: the overlapping factors of feature strings:

$$OverlapFactor_i = \frac{LCC(Sour, Dest)}{length_i}$$

$(i = Sour, Dest)$.

Definition 4: When $OverlapFactor_{Sour} \geq \alpha$ *and* $OverlapFactor_{Dest} \geq \beta$ $(\alpha \leq 1, \beta \leq 1, \beta \leq \alpha)$ is true, it means the Chinese web pages represented by $Sour$ and $Dest$ are duplicate pages. The values of α are as follows:

$$\alpha = \begin{cases} \alpha & host(a) \neq host(b) \\ \alpha + \varepsilon & host(a) = host(b) \end{cases}.$$

The $host(page)$ expresses web site domain name, and $\alpha + \varepsilon \leq 1, 0 < \varepsilon < 1$.

The $OverlapFactor_{Sour} \geq \alpha$ defines that the number of repeat characters must reach proportions α in $Sour$. But when the length difference is too large between $Sour$ and $Dest$, the expression $OverlapFactor_{Sour} \geq \alpha$ may be invalid. For example, when the length of $Sour$ is 1, the length of $Dest$ is 10 and $LCC(Sour, Dest)$ is 9, the value of $OverlapFactor_{Sour}$ is 1. In order to prevent this phenomenon, it introduces judging condition $OverlapFactor_{Dest} \geq \beta$. This condition limits the number of repeat characters must be reached proportions β in $Dest$. Generally, the templates of the sibling pages are same under a specific web site. When two web pages come from the same domain web site, their feature string contains same noises which come from copyright information or other parts. Through segmenting the value of α to artificially increase the value of α, it can avoids miscalculation of the same noise but the different content.

VI. Experiments and Analysis

This research collects 998 web pages randomly from the Internet and identifies duplicate pages artificially.

The correctness standard of algorithm evaluation adopts recall and precision. It is defined as follows:

$$Recall = \frac{the\ number\ of\ correctly\ identify\ pages}{the\ number\ of\ duplicate\ pages}$$

$$Precision = \frac{the\ number\ of\ correctly\ identify\ pages}{the\ number\ of\ identify\ pages}$$

The recall and precision on experimental data set is

shown in table 1. The result shows this algorithm is available.

TABLE I

THE STATISTICS OF RECALL AND PRECISION

Recall	Precision
96.7%	97.8%

The total consuming time of key steps in web page duplicate is shown in table 2. The edit distance solving is the main cost in the process of comparing feature strings. This paper adopts traditional dynamic programming algorithm which the time complexity is $O(m \times n)$. To improve the efficiency of this algorithm, it can use improved edit distance for solving scheme; or use map/reduce model in clusters for simultaneously calculating the edit distance of multiple feature strings.

TABLE II

THE STATISTICS OF EFFICIENCY

The time of extraction(s)	The time of comparison(s)
0.29	5.34

The experiment program is written in python. The coefficients in experiment are shown in table 3:

TABLE III

THE ILLUSTRATE OF COEFFICIENTS VALUE

a	β	ε
0.6	0.8	0.2

VII. CONCLUSION

This paper achieves fuzzy match of feature strings under the influence of noise by introducing the concept of the edit distance, taking the antisense concept of edit distance to get the concept of the largest common number of characters, and limiting proportion of the public number of characters in its original feature string. With the help of the text content of a Web page and Web page layout features, this paper selects the period as the tag of characters in feature string to make the content accounting for a large proportion in feature string as much as possible. This article avoids extraction of the Web page content by the above two points. This algorithm has simple structure with very strong parallelism, in cluster environment [16] can speed up the algorithm efficiency; with the help of psychological trend of network[17] and the Chinese word cloud technologies[18], it lays the groundwork for the subsequent processing of mass information, such as, weighting links[19], web page classification[20], and ranking web documents[21].

When there is no period or there are all pictures in article, this algorithm is invalid, because the length of feature string is insufficient or the characters of feature string are all noise.

VIII. ACKNOWLEDGEMENT

This work was supported by 2012 China National Social Science Fund (No. 12XSH019), 2011 Soft Science Plan of the Sichuan Provincial Department of Science and Technology (2011ZR0058) and 2012 Humanities and Social Sciences Planning Fund of China Ministry of Education (No. 12YJA190009).

REFERENCES:

[1] Broder A Z,Glassman S C,Manasse M S, Syntactic clustering of the Web. The Sixth International Conference On World Wide Web, 1997.

[2] Broder A Z, Identifying and Filtering Near-Duplicate Documents. Lecture Notes in Computer Science, 2000. 1848: p. 1-10.

[3] Shivakumar N, Garcia-Molina H., Finding near-replicas of documents on the web. The World Wide Web And Databases, 1999, 1590: p. 204-212.

[4] Chowdhury A, Frieder O, Grossman D, Collection Statistics for Fast Duplicate Document Detection. ACM Transactions on Information Systems, 2002, 20(2): p. 171-191.

[5] Kołcz A, Chowdhury A, Lexicon randomization for near-duplicate detection with I-Match. Supercomputer, 2008, 45: p. 255-276.

[6] WU Ping bo,CHEN Qun xiu,MA Liang, The Study on Large Scale Duplicated Web Pages of Chinese Fast Deletion Algorithm Based on String of Feature Code. Journal of Chinese Information Processing, 2003, 17(2): p. 28-35.

[7] WEI Li-xia,ZHENG Jia-heng, Detection and elimination of similar Web pages based on text structure. Journal of Computer Applications, 2007, 27(11): p. 2584-2586.

[8] HUANG Ren,FENG Sheng,YANG Ji-yun,LIU Yu,AO Min, Detection and elimination of similar Web pages based on text structure and extraction of long sentences. Application Research of Computers, 2010, 27(7): p. 2489-2491.

[9] CAO Yu-Juan,NIU Zhen-Dong,ZHAO Kun,PENG Xue-Ping, Near Duplicated Web Pages Detection Based on Concept and Semantic Network. Journal of Software, 2011, 22(8): p. 1816-1826.

[10] CHNG Pengsen,AN Junxiu, The Duplicate News Web Page Detection Algorithm based on Feature Words Group. Journal of Chengdu University of Information Technology, 2012, 27(4): p. 374-379.

[11] Lopresti D P, A Comparison of Text-Based Methods for Detecting Duplication in Scanned Document Databases. Information Retrieval, 2001, 4: p. 153-173.

[12] BAO Jun-Peng, SHEN Jun-Yi, LIU Xiao-Dong, SONG Qin-Bao, A Survey on Natural Language Text Copy Detection. Journal of Software, 2003, 14(10): p. 1753-1760.

[13] Ye S, Wen J, Ma W, A systematic study on parameter correlations in large scale duplicate document detection. Knowledge and Information Systems, 2008, 14(5): p. 217-232.

[14] Li Zhiyi,Liang Shijin, National Research on Deleting Duplicated Web Pages:Status and Summary. Library and Information Service, 2011, 55(7): p. 118-121.

[15] Levenshtein VI, Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 1966, 10(8): p. 707-710.

[16] AN Jun-xiu, Research and Demonstration of Cloud Retrieval System Based on Server Clusters. Computer Science, 2010, 37(7): p. 179-182.

[17] JIN Yu-chang, QIN Qi-wen, AN Jun-xiu, Psychological Analysis of Trends Intelligent Network Model. Computer Science, 2010, 37(6): p. 273-277.

[18] An J, Research and Implementation on the Cloud of Chinese Letter. International Conference on Information Science Automation and Material System. 2011.

[19] Xinyue Liu, Hongfei Lin, Liguo Zhang, An Attractive Force Model for Weighting Links in Query-Dependant Web Page Ranking. Journal of Software, 2012, 7(1): p. 124-129.
[20] Yanjuan Li, Maozu Guo, Web Page Classification Using Relational Learning Algorithm and Unlabeled Data. Journal of Software, 2011, 6(3): p. 474-479.
[21] Xinyue Liu, Hongfei Lin, Cong Zhang, An Improved HITS Algorithm Based on Page-query Similarity and Page Popularity. Journal of Software, 2011, 7(1): p. 170-174.