Journal of Computers Vol. 29 No. 2, 2018, pp. 104-109
doi:10.3966/199115992018042902011

A Method of Detecting Approximate Repetitive News Documents

Xueping Liang[1*], Xiaojun Wen[1]

[1]School of computer engineering, Shenzhen polytechnic, Shenzhen 518055, China
xp_liang@szpt.edu.cn

Received 24 July 2017; Revised 19 September 2017; Accepted 19 October 2017

Abstract. In view of the phenomenon of too much repeated webpage on the Internet, this paper proposes an approximately duplicate webpage detection algorithm and system，which combined multi-feature fingerprint cluster detection with document similarity detection. In this scheme, the multi-feature fingerprint cluster detection is used first to ensure the precision and efficiency of the algorithm; for small portion of the document that not be recalled, approximately duplicate webpage detection algorithm is used to guarantee the recall rate. The scheme has good improvements in the aspects of precision and recall rate, and at the same time has a good balance on performance.

Keywords: approximate repetition of documents, document clusters, multi-feature fingerprint clusters

1 The Most Similar to the Existing Technology Solutions

1.1 Technical Background

According to statistics, repeated pages on the Internet accounted for about 30% to 45%. There are pages that are exactly the same as those caused by mirroring, and there are pages caused by only small differences, such as advertisements, counters, timestamps, and so on, which are irrelevant to the content of the search. According to the statistical report, when asked “which is the biggest problem encountered on searching and retrieving the information?”, people who choose “too much repeated information” accounted for 44.6%, ranking No. 1 [1]. Eliminating similar web pages will save network bandwidth, reduce storage space, improve the quality of the index, that is, improve the efficiency and quality of the query service, while reducing the burden on the remote server where the page is located.

In the news document, due to the reprint between the various network media, the exchange of manuscripts, the use of prescribed documents in certain major events, etc., or the same news events reported little gap, so the rate of news document repeated is higher, about 60% to 80%. In the news search scenario, the user demand for eliminating similar documents is also higher, and the existing algorithm in terms of accuracy and recall rate cannot meet the requirements.

1.2 Technical Solutions of the Prior Art

(1) The traditional method of detecting the approximate document: directly compare the two text similarity, mostly after the text word segmentation, into the eigenvector distance measurement, such as the common Euclidean distance, Hamming distance or cosine angle and so on.

(2) Based on the signature of the web page: from the text of the web page, extract a small amount of information to form a signature. In classification, the pages are replaced by signature to determine whether the corresponding web content is repetitive. As the punctuation mark occurs in most of the text in the page, this method takes a fixed length of words in two sides around the period punctuation as a

[*] Corresponding Author

signature to uniquely identify the page.

(3) Simhash algorithm [2-3]: (2007 years of the paper "detected near-duplicates for web crawling"), from the massive text, fast search simhash collection which are less than *k*-bit difference with known simhash. Here each text can be represented by a simhash value, which is 64bit long. Similar text has similar simhash. In this paper *k* value is recommended to be 3.

(4) Shingling [4-5] algorithm: first with the concept of mathematics strictly define what is "roughly the same": the similarity between two documents *A* and *B* is a number between 0 and 1. So if this the number close to 1, then the two documents is "roughly the same". The definition of inclusion degree is the same. To calculate the similarity and the degree of inclusion between the two documents, only hundreds of bytes of sketch on the documents are needed. Throughout the web application, the shingling algorithm will cluster similar documents.

(5) Algorithm for combining signatures and LCS [6-7]: (1) Segmenting a web page document set into similar subsets of documents using signatures, calculating LCS only in each possible similar document subset; (2) First filter the web document to generate a document filtering framework, then performs LCS calculations on the frame instead of the original documents; and (3) calculates the LCS and selects its trusted part (called TLCS) to calculate the similarity.

1.3 The Shortcomings of the Prior Art and the Issues that Will be Addressed in this Paper

(1) One of the biggest drawbacks of the signature checking method based on signatures is that it cannot be extended to massive data and linearly compare-time complexity.

(2) The time complexity of the method based on signature is O(n), and the complexity of the algorithm is high for large number of pages. At the same time, the pattern matching is an exact match, the resistance to the page noise is poor, and recall rate is low.

(3) The shortcomings of the simhash [8-9] algorithm are as obvious as the advantages, there are two main points: for the short text, *k* value is very sensitive; the other is because the algorithm uses space for time, system memory consumption is high. The accuracy and recall rate of this method are both about 80%, which cannot achieve the desired requirements.

(4) shingling [10] algorithm: Sketch computational efficiency is relatively high, the time is a linear relationship with the size of the document. The data structure and algorithm of the design are very limited by the amount of data entered. The algorithm needs 24G when each document has an 800-bytes sketch. Precision and recall rate are not up to the ideal state.

(5) in this algorithmthe documents will be divided into a similar document subset using fingerprint, which is too dependent on the extraction of feature fingerprints. If the two similar documents feature fingerprint extraction is not the same, these two documents cannot be confirmed is similar to duplicate documents, which affects the recall rate; at the same time, the feature fingerprint is only used for segmentation of documents, not multi-feature fingerprint to determine whether there is a similar document repeat documents, which affects the efficiency of the program.

2 The Technical Program of this Detailed Description

In this paper, a multi-feature fingerprint cluster detection and document similarity detection is combined in the approximate repeat page detection algorithm and implementation system. In this scheme, multi-feature fingerprint cluster detection is used to ensure the precision and efficiency of the algorithm. For the small part of the document that is not recalled, the document similarity detection algorithm is used to ensure the recall rate. The scheme has good improvements in the aspects of precision and recall rate, and at the same time have a good balance on performance. The program structure is shown in Fig. 1.

The document repetitive detection system in this document consists of a feature extraction device, a selectionor, a comparison device, a Voter, an Updater, a fingerprint mapping system (FMS) and Document Mapping System (DMS). Given a new web page document *D*, the process is as follows:

(1) Feature fingerprint extraction device (Extractor) extract a number of features of document *D*, generate feature fingerprint.

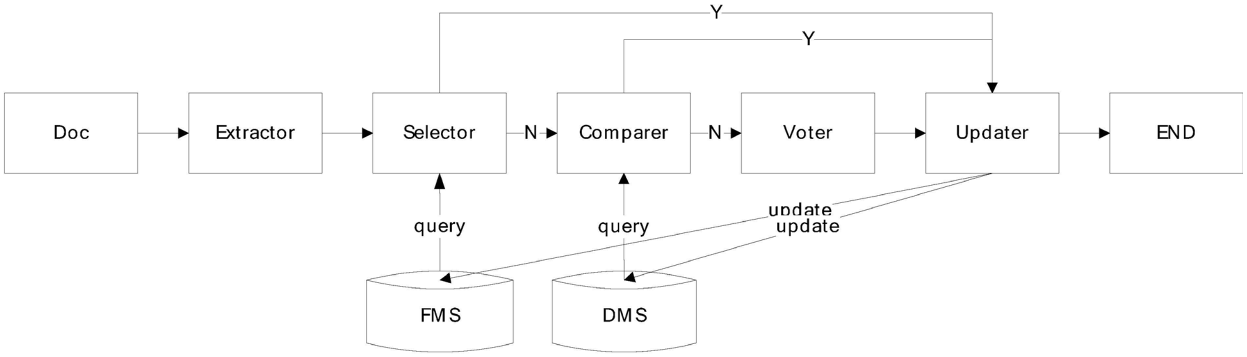

Fig. 1.

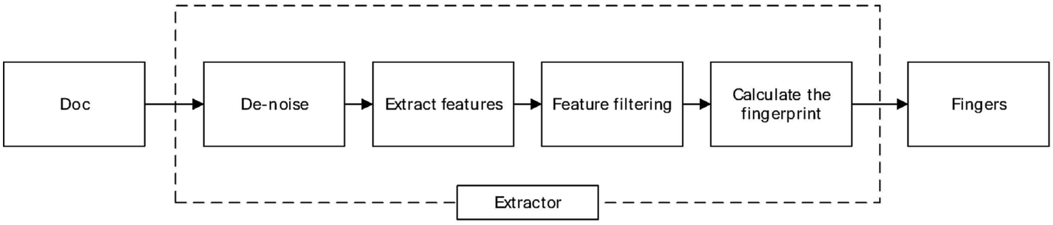

Fig. 2.

First, it de-noise the document *D*, clean up the document characters, paragraphs (remove some interference characters), transform Chinese character (full-angle to half-angle), and only retain the text and title for the need to extract feature. The contents of the text are segmented by paragraph identification to find the longest *N*-1 paragraphs. The *N*-1 paragraphs are divided into sentences according to the punctuation marks to find the longest sentence in each paragraph. The *N*-1 longest sentence and the title together form the *N* features of the document (for example, *N* is 5). Then features are de-noised, such as removing the punctuation, etc. If the length of the document feature is less than a certain value, such as 10, then the whole feature is discarded. Finally the number of features of the document is *M* (*M* >= 1, *M* <= *N*) This *M* features are calculated to generate *M* feature fingerprints.

The mathematical model for extracting the longest sentence as a feature of the document is as follows. If *S* represents a series of words w1, w2, ..., wn, in other words, *S* can represent a meaningful sentence consisting of a series of words that are rehearsed in a particular order. Now, if you want to know the possibility of *S* in the text, with P (*S*) to represent, can be expanded to:

$$P(S) = P(w_1)P(w_2|w_1)P(w_3|\ w_1\ w_2)\ldots P(w_n|w_1\ w_2\ldots w_{n-1}), \qquad \textbf{(1)}$$

For the Markov hypothesis, P (S) becomes:

$$P(S) = P(w_1)P(w_2|w_1)P(w_3|w_2)\ldots P(w_i|w_{i-1}). \qquad \textbf{(2)}$$

Thus, for a document set with *D* documents, the probability that a sentence belongs to a document can be approximated as: P (*F*) = 1 / (*D* * P (*S*)). *D* is a constant, so the longer the sentence, the lower the probability of the sentence, the greater the probability that the sentence will represent the document. So the longest sentence in the longest paragraph is the most likely to belong to the cluster of sentences extracted as a feature.

(2) Cluster selection device (Selector) use fingerprints to strictly detect whether the document *D* belongs to an existing document cluster.

The Selector determines whether the *M* fingerprints of the document *D* appear in the Fingerprint Mapping System (FMS) and detect the presence of an approximately duplicate document based on the presence of *M* fingerprints in the FMS. If there are *T* (*T* >= 1) fingerprints in the fingerprint mapping system, and if the *T* fingerprints are accompanied by *X* (such as *X* >= 2) fingerprints pointing to the same document cluster *C*, then think that document *D* belongs to document cluster *C* (this process should make the detection precision more than 99%).

The value of X depends on the requirement of precision. Suppose the precision rate of single fingerprint is P(f), the error rate is 1 - P (f). The error rate of X fingerprints $(1 - P(f))^X$, and the precision rate is $1-(1 - P(f))^X$. So in the cluster selection device (Selector), if the single fingerprint (X = 1) the precision rate is 90%, then when X = 2 the precision rate is 99 %. If matched, turn to step (5); otherwise go to (3).

(3) Document comparison device (Comparer) compares the document with the recent T-day documents, according to the similarity of the document to detect whether the document D is an existing document cluster.

News document is new and relates to recent time, which is generated with the occurrence of the event, is stopped with the end of the event. For the news document, 99% of its approximate documents is within the last few days (3-5 days). So you can find a news document whether there is a similar news, only to find the last few days on it, which greatly improves the efficiency of the calculation.

The document comparison device (Comparer) compares the new document and the similarity of the document in the last T days. Since the number of documents in the last T days is large, the titles of the documents for the last N days is segmented (Chinese Word Segmentation), for example, "I love Beijing" is segmented into "I | love | Beijing", and the stop words are discarded. Then a <word/document> inverted table is set up for each word in the titles. So in the actual similarity comparison of the new document with the recent T-day documents, first of all, segment the title of the new document, search the segmented words in the <words / documents> inverted table to find the M(such as 50) documents which hit the largest number in inverted table. Then do the similarity comparison. The comparison algorithm can be LCS (the longest common string), COS (cosine similarity comparison), etc. If the maximum similarity value x is greater than a certain threshold k (if k is 0.7), then think that the document D and the document with the similarity value x belong to the same cluster. So we detect the new document as the approximate repeat news, and find the document cluster C, turn to step (5).

(4) Cluster election device (Voter): document D does not belong to any existing document cluster, self-contained.

In case document D is self-contained, the cluster election device (Voter) generates the cluster head of the document cluster according to the election algorithm. The fingerprint of the title can be used as the value of the cluster head, or you can also take any feature fingerprints for the cluster head value.

(5) Cluster information update device (Updater) update the document D and its own document cluster information to the storage system.

After the document D finds its document cluster C, the cluster update device updates the feature fingerprints and cluster information to the fingerprint mapping system (FMS). In the process of updating, there are M(M >= 1) feature fingerprints in the document, and the feature fingerprint F in FMS appears in three cases: did not appear in the FMS; appeared in the FMS, but it belongs to the document cluster is not C; appeared in the FMS, while it belongs to the cluster is C. So a weight value is need to indicate whether the document feature fingerprint F should binds to its cluster C or not. Fingerprint F in the FMS appear in the situation are:

(A) Did not appear in the FMS: the weight is initialized to a0.

(B) Occurs in the FMS, but it belongs to the document cluster that is not C: reduce weight, when the weight value is reduced to less than a certain threshold, the lifting of the relationship between F and C.

(C) Appears in the FMS, and the document cluster belongs to the C: increase weight, indicating that the relationship is further determined.

Cluster information update device (Updater) will also updates document D and cluster headers to the document mapping system (DMS). As described in the document comparison device (Comparer), the title of document D is segmented, and then create an inverted table of <words / documents> for each word.

3 For 2 of the Technical Program, Whether There Are Other Alternative

(1) The extraction of the feature fingerprint extraction device (Extractor) can have a variety of algorithms, take the full text as a feature, take the paragraph as a feature, take fixed length words around the period punctuation as a feature, slice the document as a feature. Paragraph as a feature is with the highest precision, but the recall rate is low; and for fixed length words around the period punctuation as a feature, the fixed length parameter selection is a problem.

(2) In the cluster selection device (Selector) to detect whether the document *D* is an existing document cluster, while there are *X* fingerprints pointing to the same document cluster, where *X* can be changed according to actual needs, can be 1 to *M* (*M* for the fingerprint Number).

(3) Document comparison device (Comparer) compares the similarity of new documents with the recent *T*-day documents. You can directly compare it with each document in the last T days, or you can use a small retrieval system which implements text match. The scheme is just a way to reduce the candidate set.

(4) The relationship between the feature fingerprint F and the cluster *C* pointed to by the cluster information update device (Updater) can also be time-dependent. When F lasts the time point to the cluster *C* exceeds a certain threshold, the relationship can be canceled.

4 Experimental Results and Comparative Analysis

In order to evaluate the correctness and efficiency of this algorithm, a series of experiments are designed.

Correctness is the life of the algorithm; here are two evaluation criteria: repeat page recall rate (Recall) and to deduplication (Precision), defined as follows:

Recall = number of webpages that are correctly de-duplicated/number of duplicated webpages.

Precision = number of webpages that are correctly de-duplicated/number of all webpages that are de-duplicated.

In order to detect the performance of the algorithm, we applied this algorithm in news search scenario and selected 20 queries in three fields: business, biology and computer. By searching the keywords in the search engine, we selected 1545 pages with the same or similar content in each group. And insert these approximate pages into an existing set of documents (including 928,518 pages). And run simhash and the algorithm for similar web detection. The test results are shown in Fig. 3.

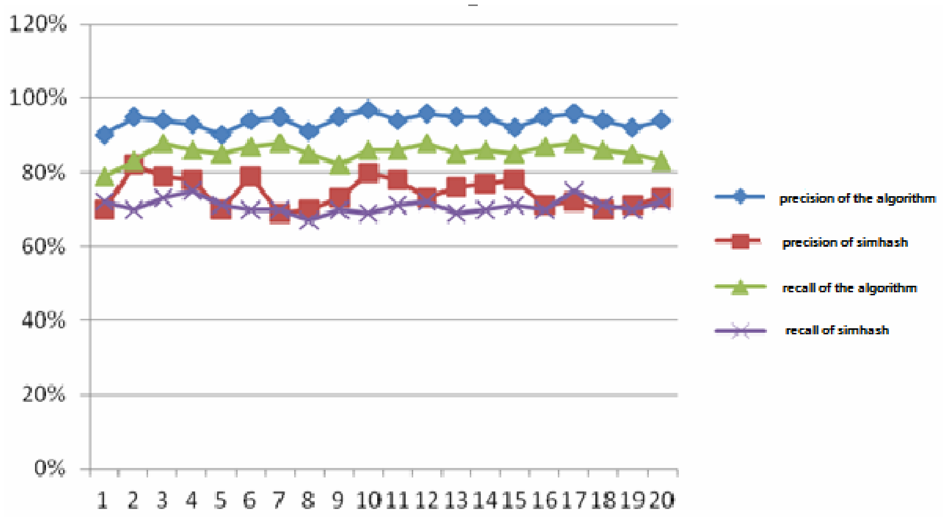

Fig. 3. Precision and recall comparison with simhash algorithm

The main factor that affects the precision of de-duplicating the web pages is the web page noise. The algorithm achieves good results with a precision rate of > 95% and a recall rate of > 85%. As the algorithm has approximate linear time and space efficiency, the experiment proves that it is suitable for large-scale Chinese web page de-duplication. The research results of this paper have been successfully applied to practical engineering projects.

5 Conclusion

In this algorithm, the multi-feature fingerprint cluster detection is used first to ensure the precision and efficiency of the algorithm; for small portion of the document that not be recalled, approximately duplicate webpage detection algorithm is used to guarantee the recall rate. The algorithm is successfully applied in news search scenario, and is proved to have good improvements in the aspects of precision and recall rate, and at the same time with a good balance on performance.

The technical key points of this article are as follow:

(1) Document feature extraction method.

(2) The combination of multi-feature fingerprint cluster detection and document similarity comparison as the second detection to solve the problem.

(3) Multi-feature fingerprint to determine the document cluster that a document belongs to, and the update method of the feature fingerprints and the document cluster that they belongs to.

References

[1] China Internet Network Information Center (Ed.), 16th Statistical Report on Internet Development in China, China Internet Network Information Center, Beijing, 2005.

[2] M.S. Charikar, Similarity estimation techniques from rounding algorithms, in: Proc. 34th Annual ACM Symposium on Theory of Computing, 2002.

[3] N. Rezaeian, G.M. Novikova, Detecting near-duplicates in Russian documents through using fingerprint algorithm simhas, Procedia Computer Science 103(2017) 421-425.

[4] A.Z. Broder, S.C. Glassman, M.S. Manasse, G. Zweig, Syntactic clustering of the web. Computer Networks 29(1157-1166)(1997) 8-13.

[5] C. Ma, Web page search algorithm Shingling and Simhash research, Computer and Digital Engineering 37(1)(2009) 15-17.

[6] L. Huang, L. Wang, A similar web page detection algorithm based on LCS, 2009.

[7] L. Chen, G.-S. Wu, J. Li, Duplicate detection for Chinese texts based on semantic fingerprint and LCS, Computer Engineering & Software 11(2014).

[8] G. Zhang, A fast search optimization method based on simhash for massive similar documents, Command Information System and Technology 2(2015) 61-65.

[9] B. Dong, Approximate text detection based on multi-simhash fingerprint, Small Microcomputer System, 2011.

[10] C. Ma, X. Mao, Research on near-duplicate detection algorithm shingling and simhash, Computer and Digital Engineering 1(2009).