

An efficient method to detect duplicates of Web documents with the use of inverted index

Sergey Ilyinsky, Maxim Kuzmin, Alexander Melkov, Ilya Segalovich

Abstract

The growth of the Internet challenges Internet Search Engines as more copies of Web documents flood over search results making them less relevant to users. A method of "descriptive words" for definition of near-duplicates of documents is suggested. It is based on the choice of N words from the index to determine a "signature" of a document and can be applied to any search engine based on the inverted index. It is compared with the method based on "shingles". At a practically equal accuracy of algorithms, this method is more efficient in the presence of inverted index.

Introduction

The growth of the Internet challenges Search Engines as more copies of Web documents flood over search results making them less relevant to users. The nature of copies is wide. The same document served from the same server may differ because of technical reasons – different character sets, formats, and inclusions of advertisement or current date. On the other hand similar documents are massively generated by database servers e.g. messages in web forums, product pages in e-shops, etc.

The problem of detecting near-duplicates has been explored in the works [1,2,4,5] of different research groups. The decision suggested there could generally be described as follows. Scan each document, compute hash values ("fingerprints") at every, possibly overlapped, fragment of the document ("a shingle"), then sample the obtained set with some reasonable criteria and use this reduced set of checksums ("a sketch") in all the following set operations. The fragments could be sentences [2] or sequences of bytes [1], consonants [4], or words [5]. Fixed and variable size sketches were suggested: a number of smallest checksums and all the checksums divisible by some integer. In the work [3] a method of choosing a set of words and an improvement to vector space model for pair-wise document comparison were suggested.

The method of "shingles" decreases the set size for each document and allows to simplifying the clustering task. In the same time the clustering of documents remains computationally hard and as a rule requires inversion of document-to-shingle relation. We implement a digital signature of the document that is probabilistically stable to small changes of the document thus the clustering procedures become redundant. All currently implemented web search engines already have a database or "inverted index" of all the words and corresponding document identifiers. The typical availability of global word

statistics enables a reasonable choice of words for accurate and efficient decision of the problem. While corresponding approach with shingles seems to be much more complicated because of the huge shingle set cardinality.

Problem statement

Let us have a set of documents. A document is seen as a sequence of words. Consequently we deal only with lexical equivalency. We suggest an algorithm that constructs a "set of duplicates". The accuracy of the algorithm can be expressed in terms of probabilities of two kinds of errors. We name "alpha-error" a situation when the algorithm has not determined a duplicate, and "beta-error" the case when algorithm has offered a false duplicate. Apparently there is a trade-off between two kinds of errors; therefore there is no ideal algorithm.

Therefore it is necessary to divide our task into two parts – the first is to construct a "set of duplicates", and the second is to check it.

Method of "descriptive words"

In the beginning it is necessary to choose some set of N words, which we name "descriptive set"; the choice of the words will be discussed later.

For each word let us fix a threshold frequency b_i and for each document compute a vector where the i -th component of the vector is set to 1 if the value of relative frequency of the i -th word from the "descriptive set" in this document is greater than the selected threshold frequency, 0 – otherwise. This binary vector is regarded as a fuzzy digital signature of the document. Each vector unambiguously determines a class of similar documents. Therefore its unique identifier can substitute a corresponding vector. Vectors with 3 or less words above threshold are removed from further consideration.

Let us formulate the basic criteria for the choice of words

1. A set of words should cover the maximal possible amount of documents
2. The "quality" of a word in the sense described below should be the highest
3. The number of words in the set should be minimal

The "quality" of a word can be determined as a relative stability of the corresponding component of the vector to small changes of a document. It means that for a "good" word i the probability of transition (let us call it λ_i) through a threshold value is minimal for small changes of a document. Obviously the probability of the change of a vector for a document is

$$P = 1 - \prod_i (1 - \lambda_i) = \sum_i \lambda_i + o(\lambda_{\max})$$

The threshold b_i we choose satisfies following criteria – the value of λ_i is minimal under the condition that both fractions (above and below threshold) are not too small. This condition is required to assure that almost all documents have non-zero vectors

The optimal number of words (N) in the "descriptive set" we determined experimentally.

Experiments

We used the normalized Levenshtein distance to estimate the similarity between two documents. The documents were considered practically equivalent if the difference between them didn't exceed 8%. This value was obtained in experiments when 6 experts made 300 assertions on document similarity: we asked experts to tell which pairs are identical from the point of view of a search engine user. The texts were preliminarily stripped from HTML-markup. Our metrics currently doesn't take into account the importance of unmatched words, i.e. word frequency, that seem to be significant in case of e-shops and other database generated documents.

We took 60 million documents from the crawl of the Yandex [6]. Then we constructed clusters of duplicates for several sizes of "descriptive set" and checked each cluster for "practical equivalence" (8%) as described above. The considered set sizes were from 1600 to 3000 words. Decrease of the "descriptive set" size decreases alpha-error and increases beta-error. At some point the number of "correctly determined" duplicates stopped its growth while the number of false duplicates stayed low. The chosen set size was 2000 words.

In the following table we show the comparison of our method to the method of "shingles", on the same test. "Shingles" were calculated according to [5] with 10-words in a shingle, 48-bits checksum, modulus 16, and the similarity threshold at 7/8 of the shingles set.

Table 1. Comparisons of the method of descriptive words and the method of "shingles"

	Number of suggested duplicates	Difference < 8%	Difference < 15%	Difference < 30%	Computing time
Method of "descriptive words"	22 millions	61%	76%	91%	~1.5 hour
Method of "shingles"	19 millions	66%	80%	94%	~3.5 hour

Conclusion

The algorithm uses only inverted index that is available in the most of search engines. It is very fast. It is easily upgraded to incremental version because the "fuzzy digital signature" is a global characteristic of a document and the database doesn't require re-clustering on each incremental crawl. At a practically identical accuracy of algorithms, this method is more efficient in the presence of inverted index.

References

1. U. Manber. Finding similar files in a large file system. Proceedings of the 1994 USENIX Conference, pp. 1-10, January 1994.
2. S. Brin, J. Davis, H. Garcia-Molina. Copy Detection Mechanisms for Digital Documents. Proceedings of the ACM SIGMOD Annual Conference, San Francisco, CA, May 1995.
3. N. Shivakumar, H. Garcia-Molina. SCAM: A Copy Detection Mechanism for Digital Documents. Proceedings of the 2nd International Conference on Theory and Practice of Digital Libraries, Austin, Texas, 1995.
4. Nevin Heintze. Scalable Document Fingerprinting. Proceedings of the Second USENIX Workshop on Electronic Commerce, Oakland, California, November 18-21, 1996.
5. Andrei. Z. Broder, Steven. C. Glassman, and Mark. S. Manasse. Syntactic Clustering of the Web. In Proceedings of the Sixth World Wide Web Conference, 1997.
6. <http://www.yandex.ru/>