http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (12): 2465-2469, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.2465.2469

A Chunk-based Copy Detection Approach for Multimedia Documents

[1,2]Li Guo, [3]Bo Jin and [1]Degen Huang
[1]School of Computer Science and Technology, Dalian University of Technology,
Dalian 116024, China
[2]School of Information Science and Engineering, Dalian Polytechnic University,
Dalian 116034, China
[3]School of Innovation Experiment, Dalian University of Technology,
Dalian 116024, China

Abstract: Copy detection is important to both intellectual property protection and information retrieval. Prior researches on copy detection (e.g., hash-based fingerprinting algorithm, etc.) concentrated on document-level copy detection. These researches have difficulty to detect complicated multimedia documents. In this study, a chunk-based copy detection approach is discussed. And a Fingerprint-based Heuristic Algorithm (FHA) is proposed based on fixed-length chunks and overlap of chunks. With experiments and results, the proposed approach can deal with both total copy detection and partial copy detection of multimedia documents.

Key words: Copy detection, multimedia documents, fingerprint-based heuristic algorithm

INTRODUCTION

Currently, Internet and digital libraries provide vast amounts of digitized information on-line. Thus, it is easy to copy someone else's work and submit it as someone's own. One can copy documents with illegal purposes, that is the plagiarism. Internet makes it easier to plagiarize other's ideas. Thus, it is a great challenge to detect and prevent plagiarism.

Nowadays, there are many content-based document copy detection systems. For example, COPS (Xiao *et al.*, 2011), Koala (Urvoy *et al.*, 2008), SCAM (Stajano and Wilson, 2011), Shingling (Elmagarmid *et al.*, 2007), I-Match (Puppin *et al.*, 2010), etc. Nowadays, there are huge amount of multimedia documents on the Internet. Multimedia documents are much more complicated. Thus, many techniques used in the above systems for normal documents are not suitable for multimedia documents, such as stemming and anchor strategy (Potthast *et al.*, 2011).

Moreover, the current copy detection systems mainly focus on document-level copy detection and struggle to measure the amount of overlap in document-level. But these systems are not always reliable for total copy detection of documents with varying size, not to mention partial copy detection. In this study, a Fingerprinting-based Heuristic Algorithm (FHA) is presented to calculate document-level and passage-level similarity for multimedia documents. Firstly, the documents are split into fixed-length chunks. Then, the overlap of these chunks is used to deal with total and partial copy detection.

CHUNK-BASED COPY DETECTION APPROACH

Generally, copy detection is to detect whether a document is the copy of other documents. In this study, a document is regarded as a set of tokens. These tokens can be characters, words, sentences, passages, etc., Let A and B be the chunk sets of two documents. Then, the relations of them include: (1) If A≈B, A is a copy of B. This is the total copy; (2) If A∈B and A≠B, A is the subset of B. This is the partial copy.

In this study, the similarity of two documents is measured in two levels: Document-level and passage-level. From the distinct set of chunks of two documents, the absolute similarity between documents is calculated using the concept of intersection of sets. Then it is possible to verify how much of candidate document is contained in another plagiarized document, as shown:

$$\text{SimD}(D_1, D_2) = \text{Max}\left(\frac{|\text{CK}(D_1) \cap \text{CK}(D_2)|}{|\text{CK}(D_1)|}, \frac{|\text{CK}(D_1) \cap \text{CK}(D_2)|}{|\text{CK}(D_2)|}\right) \quad (1)$$

where, CK (D) represents the set of whole chunks of document D.

Corresponding Author: Li Guo, School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China

```
Initialization:
    DocSet←{$d_1$, $d_2$, ..., $d_n$}
    D_Threshold←0.6
    C_Threshold←0.35
Main:
While (DocSet≠Ø) do
    $d_i$, $d_j$←pop (DocSet)
    If (SimD($d_i$, $d_j$)>D_Threshold) then
        CopyFlag←1
    EndIf
    ChunksSet←{$c_1$,$c_2$,..., $c_m$}
    While (ChunksSet≠Ø) do
        $w_i$, $w_k$←pop (WindowsSet)
        If (SimC($c_i$, $c_k$)>C_Threshold) then
            SimC_Sum←SimC_Sum+1
        EndIf
        If (SimC_Sum/ChunkCount>D_Threshold) Then
            CopyFlag←1
        Else
            CopyFlag←0
        EndIf
    EndWhile
EndWhile
```

Fig. 1: FHA: Fingerprinting-based heuristic algorithm

In the same way, the concept of intersection and union of set is used to calculate the similarity of chunks in documents, as shown:

$$\mathrm{SimC}(C_1, C_2) = \frac{|\mathrm{FP}(C_1) \cap \mathrm{FP}(C_2)|}{|\mathrm{FP}(C_1) \cup \mathrm{FP}(C_2)|} \tag{2}$$

where, FP (C) represents the set of whole fingerprints of chunk C.

In practice, the numerator represents the total number of fingerprints in chunks of the plagiarized document and candidate document and the denominator represents the sum number of fingerprints occurring simultaneously in chunks of two documents plus the number of fingerprints of chunks that occurs in each of the documents that do not occur in the other one.

Figure 1 showed the instantiation FHA of the copy detection approach. As in the untimed case the algorithm is based on a document-set, DocSet, containing all the documents and a chunk-set, ChunkSet, containing all the chunks of a document.

Document parsing: In the proposed copy detection approach, each document is first passed through a stopword-removal process, which removes all the stopwords. Stopwords in a document should first be removed since stopwords, such as articles, conjunctions, prepositions, punctuation marks, numbers and non-alphabetic characters, often do not play a significant role in document. This process reduces the size of a document for comparison and subsequently the complexity on copy detection.

Prior researches on copy-detection prototype systems, e.g., SCAM, Koala and Shingling, have the common approaches that they split the document into chunks. It is showed by Bar-Yossef *et al.* (2009) that the length of chunks should be somewhere between 40 and 60 characters for plagiarism-detection purposes. But this selection strategy of chunk does not fit to the copy detection for complicated multimedia document.

In this study, the document is split into fixed-length chunks. Some kind of culling strategy is used to select representative chunks. A short chunk is not very representative of a text. The fact that two files share a short chunk does not lead us to suspect that they share ancestry. In contrast, very long chunks are very representative, but unless a plagiarizer is quite lazy, it is unlikely that a copy will retain a long section of text. It's better to retain similar chunks for any file.

Chunk parsing: It has been experimented with two culling methods for chunks. For candidate documents, they should be stored in the database, so non-overlap culling method of chunk is used to save space of database. For plagiarized document, they should be detected carefully, so overlap culling method of chunk is used to get higher precision.

If the proposed approach is widely used and the size of chunks is published, plagiarizer can managed to elude copy detection. It should set variable parameters to change the size and location of chunks. Then the location of chunks is:

$$Q = \theta l + \sum W_i^{Q} (\theta \in [0,1]) \tag{3}$$

where, l is the size of chunk, θ is variable size parameter of chunk. l and θ should be changed aperiodically to assure the algorithm security.

Fingerprint parsing: In this study, there are different methods to measure document-level and passage-level fingerprints. Full fingerprint is used for document-level, overlapped fingerprint for passage-level. For chunk is fixed-length, its fingerprint resolution is also fixed. Let g (g>1 word) be the fingerprint granularity, SizeC be the chunk size, n be the overlap rate. For document-level, the location of fingerprints in chunk is: 0, 1, 2, …, SizeC. For passage-level, the location of fingerprints in chunk is: 0, g-n, 2(g-n), …, SizeC.

In fingerprinting, Karp-Rabin hash algorithm Puppin *et al.* (2010) is adopted to generate fingerprints. And a 64-bit hash value is produced to ensure better distribution of hashing.

RESULTS

The collection used in the experiments was composed by about 15,000 documents. These documents is downloaded from Internet and used for research purposes only.

Evaluation: Recall and precision are widely used as metrics for information retrieval. For copy detection, a better metrics is the combination of Highest False Match (HFM) and separation (Sep.) (Potthast *et al.*, 2011). HFM is the highest percentage given to an incorrect result. The Sep. is the difference between the lowest correct result and the HFM. Their ratio (Sep./HFM) can be used to verify the differentiability of algorithm. All experiment results are averages over 10 queries and recall is measured at 20 results retrieved.

Threshold value: Firstly, it should confirm suitable threshold value for FHA. Threshold value depends on chunk size and overlap rate. As shown in Fig. 2, when the expected precision is 1, the proper threshold value should be about 0.3 to 0.35 from the results.

Chunk culling policy: In this experiment, the document is spilt into chunks of different sizes, e.g., 64, 128, 256 and 512 words. The fingerprint granularity is set to be 3 words, the overlap rate of chunks to be 0.5 and the overlap rate of fingerprints to be 1. As shown in Table 1, the best result appears when the size of chunk is 256 words. The size of 128 words also produced an acceptable result. But when the size of chunks is 512 words, the differentiability plays good, but the recall is very low. When the size of chunks is 64 words, almost all of the results are good enough, but the cost time will be too large. There is an experiment for cost time as following.

Figure 3 showed that chunk size and overlap is the important influence to the efficiency of algorithm. From results it can be seen that when the overlap descends and the chunk size rises, cost time will descend. Taking range of 128-256 words as chunk size and 0.5 as overlap rate is the proper settings for the proposed algorithm.

Fingerprinting policy: The impacts of fingerprint include culling policies of fingerprints and fingerprint granularity. Nowadays, there are mainly two culling policies of fingerprint: Total subset and total fingerprint. The total fingerprint policy can get the largest amount of fingerprints, the location of fingerprints in this policy is: 0, 1, 2, ..., SizeC-g, where, SizeC is the size of chunk, g is the fingerprint granularity. The total subset policy can get the largest granularity of fingerprints, the location of fingerprints in this policy is: 0 g, 2 g, ...,SizeC-g. Too large amount of fingerprint will lead to high cost time and too large granularity of fingerprint will lead to low precision. In this study, the variable fingerprint policy is proposed to balance the two policies. Table 2 shows the comparisons of different policies.

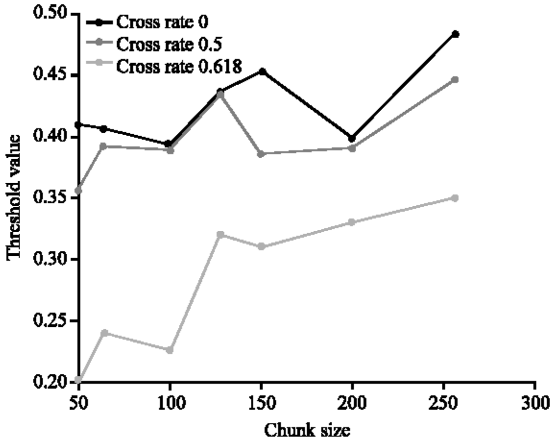

Fig. 2: Graph of average threshold value

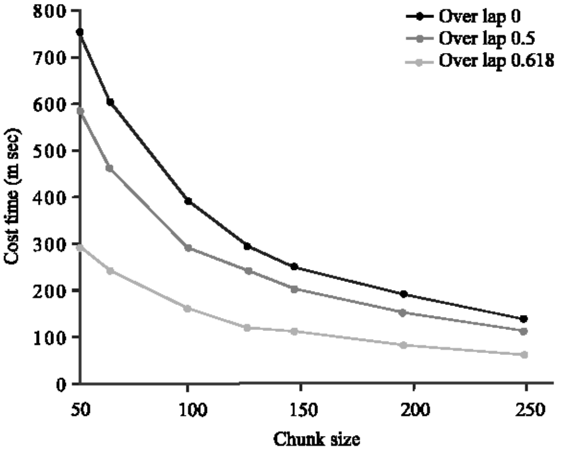

Fig. 3: Graph of average cost time according to chunk size

Table 1: Impact of changing the chunk size

Size	Precision	Recall	HFM (%)	Sep. (%)	Sep./HFM
W-64	0.96	0.965	25.03	6.03	0.24
W-128	0.83	1	13.08	7.98	0.684
W-256	0.53	1	8.23	7.42	1.042
W-512	0.31	1	5.05	5.07	1.334

HFM: Highest false match, Sep.: Separation

Table 2: Comparisons of different culling policies of fingerprint

Policy	Precision	Recall	HFM (%)	Sep. (%)	Sep./HFM
Total subset	1	0.35	9.16	6.76	0.74
Variable fingerprint	1	0.70	8.92	8.12	0.91
Total fingerprint	1	0.83	7.57	8.00	1.06

HFM: Highest false match, Sep: Separation

Table 3: Impact of changing fingerprint granularity

Gran	Precision	Recall	HFM (%)	Sep. (%)	Sep./HFM
G-1	1	0.352	43.50	20.70	0.523
G-2	1	1	16.60	34.60	2.526
G-3	0.73	1	8.23	7.42	1.041
G-4	0.41	1	5.03	7.37	1.607
G-5	0.15	0.75	3.72	-0.30	-1.389
G-6	0.13	0.67	2.33	-1.20	-1
G-8	0.2	0.17	2.67	-1.70	-0.488
G-10	0.2	0.17	0	0	0
G-20	0	0	0	0	0

HFM: Highest false match, Sep.: Separation

Table 4: Comparisons of different algorithms for total copy detection

Policy	Precision	Recall	HFM (%)	Sep. (%)	Sep./HFM
I-Match	1	0.17	0.54	0.87	1.61
*COPS	1	0.53	2.84	2.46	0.87
FHA	1	0.76	19.60	6.24	0.32

HFM: Highest false match, Sep.: Separation, *Xiao *et al.* (2011), FHA: Fingerprint-based heuristic algorithm

Table 5: Comparisons of different algorithms for partial copy detection

Policy	Precision	Recall	HFM (%)	Sep. (%)	Sep./HFM
I-Match	1	0.001	/	/	/
*COPS	0.37	0.23	3.71	-1.52	-0.41
FHA	1	0.75	9.07	7.63	0.84

HFM: Highest false match, Sep.: Separation, *Xiao *et al.* (2011), FHA: Fingerprint-based heuristic algorithm

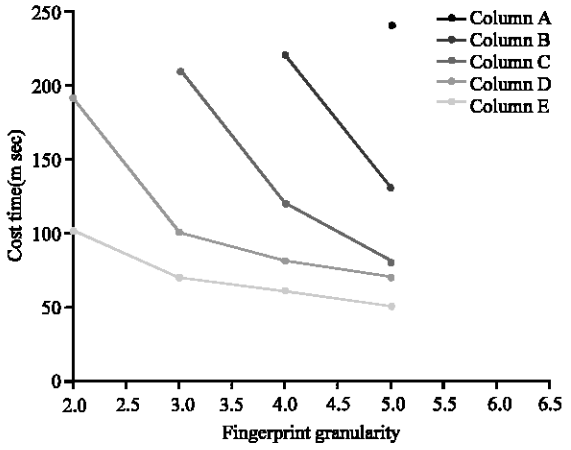

Fig. 4: Graph of average cost time according to fingerprint granularity

In the experiment, the chunks size is set to be 256 words, the overlap rate of chunks to be 0.5, the cross rate of fingerprints to be 1. The fingerprint granularity is set with different value (from 1-8 words). As shown in Table 3, the best result appears when the fingerprint granularity in the range of 2-4 words.

Figure 4 showed that when fingerprint granularity is 3 and the overlap is 1, the best result can be got. Then, the final setting of FHA can be got: Chunk size is 256 words, fingerprint granularity is 3 words, overlap rate of chunk is 0.5, overlap rate of fingerprint is 1. Then, latter experiment will use this setting.

Comparisons of different algorithms of copy detection: Table 4 and 5 showed the comparisons of different algorithms for copy detection. For total copy detection, the proposed algorithm was able to obtained 76% recall with 100% precision, as shown in Table 4. In comparison, the COPS obtained 53% recall and the I-Match obtained only 62% recall with 100% precision. The proposed algorithm also wins in the comparison of differentiability. For partial copy detection, the proposed algorithm obtained 75% recall with 100% precision, as shown in Table 5. In comparison, the COPS obtained 37% recall and the I-Match cannot deal with the partial copy detection.

CONCLUSION

In this study, an approach is presented which combine document-level and chunk-level similarity measure to detect total and partial copying. A fingerprint-based heuristic algorithm is also proposed. The algorithm adopted the fingerprinting method and use heuristic method to perform better than traditional copy detection method. From experiment, the algorithm can effectively detect copying of complicated multimedia documents with widely varied size.

ACKNOWLEDGMENTS

This study is supported by the National Natural Science Foundation of people's Republic of China (Grant Nos. 51105052).

REFERENCES

Bar-Yossef, Z., I. Keidar and U. Schonfeld, 2009. Do not Crawl in the DUST: Different URLs with similar text. ACM Trans. Web, Vol. 3.

Elmagarmid, A.K., P.G. Ipeirotis and V.S. Verykios, 2007. Duplicate record detection: A survey. IEEE Trans. Knowledge Data Eng., 19: 1-16.

Potthast, M., A. Barron-Cedeno, B. Stein and P. Rosso, 2011. Cross-language plagiarism detection. Lang. Resour. Eval., 45: 45-62.

Puppin, D., F. Silvestri, R. Perego and R.A. Baeza-Yates, 2010. Tuning the capacity of search engines: Load-driven routing and incremental caching to reduce and balance the load. ACM Trans. Inform. Syst.,

Stajano, F. and P. Wilson, 2011. Understanding scam victims: Seven principles for systems security. Commun. ACM., 54: 70-75.

Urvoy, T., E. Chauveau, P. Filoche and T. Lavergne, 2008. Tracking Web spam with HTML style similarities. ACM Trans. Web., Vol.2.

Xiao, C., W. Wang, X. Lin, J.X. Yu and G. Wang, 2011. Efficient similarity joins for near-duplicate detection. ACM Trans. Database Syst.,