International Journal of Multimedia and Ubiquitous Engineering
Vol.11, No.5 (2016), pp.209-220
http://dx.doi.org/10.14257/ijmue.2016.11.5.19

Text Recognition Algorithm Based on Text Features

De Li[1], XueZhe Jin[1] and LiHua Cui[2*]

[1]*Department of Computer Science, Yanbian University*
133002, Yanji, China
[leader1223@ybu.edu.cn, xuezhekim@gmail.com]
[2]*College of Economics and Management, Yanbian University*
133002, Yanji, China
[e-mail: 2732677@163.com]
**Corresponding author: LiHua Cui (2732677@163.com)*

Abstract

It is difficult to realize the text watermarking algorithm on natural language, and the format of text watermarking algorithm has poor robustness against format attacks. This paper presents the new text recognition algorithm based on the text feature. The words are segmented and extracted according to the text feature. The feature dimensions are reduced with the technology of LSA and stop-words database. The new similarity method is also defined to determine the threshold in order to detect the watermarking. The experimental results indicate that the proposed algorithm has better operating efficiency and stronger robustness than the previous researches. This algorithm can also handle the text document written in both Chinese and English effectively.

Keywords*: Word segmentation, frequency, text feature, similarity method, latent semantic analysis*

1. Introduction

With the development of information technology, much more official documents, papers, patents, novels and many others use the electronic form for storage, transmission and distribution in the network of e-commerce, government, *etc.* However, the electronic documents can be copied without leaving any trace, so the issues such as illegal distribution, tampering and reselling, have become more and more serious. The existing copyright protection mainly focuses on video and audio. But these are also very important issues for text documents.

The current research about text watermarking mainly focuses on text forms and text semantics. For instance, the previous watermarking algorithms of natural language are based on the structure of the document. These methods embed watermark mostly by changing the semantics of the text [1-4] or adjusting the forms of the text contents [5-7], but there are obvious deficiencies in capacity and robustness. Audio and video have enough redundant information fields, however text does not. Therefore, the watermark embedded in the text can be detected easily. In addition, the infringement behavior can be done easily the network such as copy the original parties, the shift transformation of the original work, synonym replacement and so on. So the traditional methods about text copyright are not very effective [8-12].

Because the text document has a small redundancy and limited space to embed watermarking, the traditional text watermarking methods could not solve the contradiction between robustness and imperceptibility **[**13-15**]**. In this paper, we proposed a text watermarking algorithm based on text features. In this method, we embedded the watermark without modifying anything of the original text with the text features.

ISSN: 1975-0080 IJMUE

Reference [16] reflected the information of the text, and only slight modification would not affect the text watermarking. Juanjuan Shu has proposed a method to embed the watermark combining with the natural language [17], but the robustness should be improved; Qin Si proposed a method to construct the watermarking by extracting features from the text and then compute the similarity by using the method of Levenshtein Distance [18], but features extracted from different texts are very similar sometimes. This method cannot separate them; ZUNERA J and other papers have presented the algorithm which embed watermarking by using the characteristics of English words [19-24], to extract features from the English word. This does not touch deep into the part of speech and semantic. All the above algorithms have used zero-watermarking technology that does not make any changes to the original text. In this paper, we proposed a similarity calculation method, which is used for determining the threshold value. The algorithm has reduced the dimensions of text features. The copyright attribution will be ensured by comparing the features extracted from the text with features in the IPR database.

2. Relevant Knowledge

The algorithm in this paper extracts features by statistics on word frequency and the information of the text. The structure of the zero-watermarking algorithm is completely invisible to the reader. It has solved the contradiction between robustness and invisibility basically from the root. If the watermarking is generated according to the text features, it can be robust regardless of any kind of attacks. However, the information of watermarking should be stored in a trusted third party-IPR database. The third party will verify the ownership of the copyright when the copyright dispute occurs.

Xiaolong Wang has proposed a method to compute the similarity of text with other documents [16]. In his research, first of all, some useless words of the text have removed according to stop words database, and then it segments the content of the text and calculate the frequency of every segmented word in the text at the same time. Next, the segmented words are assembled into a string, and the similarity are computed using the Levenshtein Distance shown as the equation (1).The symbol of $LD(s_1, s_2)$ represents the Levenshtein Distance between the strings $s_1$and s_2, and the symbol of $MaxLen(s_1, s_2)$ represents the bigger value of string length between the strings s_1 and s_2. The symbol of $Sim(s_1, s_2)$ is the similarity value between the strings s_1 and s_2.

$$Sim(s_1, s_2) = 1 - LD(s_1, s_2) \Big/ MaxLen(s_1, s_2) \quad (1)$$

The same steps will be done to text documents stored in the database. If the value of the similarity is higher than the threshold, it is supposed that two compared texts are same; otherwise it will be considered as a new text which will be registered in the IPR database. This method uses word frequency as the text feature. It is very effective, but the computed similarity with Levenshtein Distance has a bad effect in experiments. The method of Levenshtein Distance can get a similar value when computing the similarity of two quite different text documents. Even replacing the key word in the text or just modifying the paragraph location would have great influence on the value of similarity. So the algorithm like this has very poor robustness.

Juanjuan Shu proposed a method to compute the similarity on natural language [17]. In his research, after segmenting the text into words with preprocessing steps, the algorithm uses words with an intermediate frequency as the feature of the text. Then the algorithm calculates the similarity value by using the Jaccard model. The Jaccard model is shown as Figure1. Symbols of q, r, s and t are four state components made up of the string to be compared. All segmented words composed of a set R. The common parts of the string s_1 and the string s_2 is named set q. The symbol of s represents the set of objects that exist in the string s_1, but not in the string s_2.The symbol of r represents the set of objects that exist

in the string s_2, but not in the string s_1.The symbol of t represents the part of objects that exist neither in the string $s_1$nor in the string s_2. The similarity of this two text features is described as the equation (2).

$$sim = q / (q + r + s) \tag{2}$$

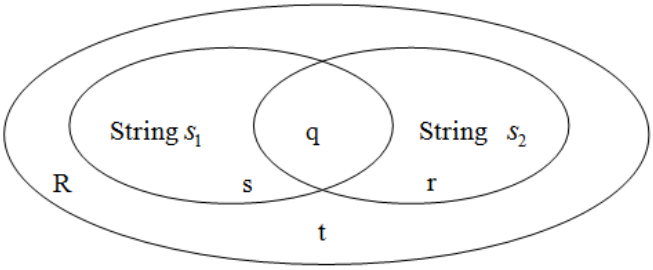


Figure 1. The Model of Jaccard

The algorithm also uses Logistic to deal with features; the process is described as the equation (3)

$$x_{n+1} = \mu x_n (1 - x_n); x_n \in [0,1], \mu \in [0,4] \tag{3}$$

This algorithm has been improved so that the robustness is stronger than the algorithm proposed by Juanjuan Shu [17], with resistance of the attack-changes of the text structure. But the algorithm based on natural language is complex in computation. In addition, it costs storage. Although it can resist some attacks of changes of the text structure, it could not be applied to large projects.

3. Identification Algorithm for Text Feature

Zero-watermarking algorithm is similar to feature-based algorithm; but zero-watermarking does not make any changes to the original content. Tampering or plagiarism is usually in connection with the modification of the format of files, the mode of expression or attack that does modify the theme of the text. In order to ensure the robustness of the watermark, this paper constructs the watermark by extracting theme features of the text document. The key to the zero-watermarking algorithm is to construct watermark with text features, but this kind of feature-based watermarking is not similar to the traditional watermarking which has specific content or meaning.

The proposed algorithm is similar to zero-watermarking; it extracts the theme features of the text content, and it does not modify the original text content. Using this method can verify the ownership of the copyright. It is protected no matter what attacks to the text document are, such as additions, deletions or composing. So the algorithm in this paper has formidable robustness. The extracted text features should be saved in the IPR database to guarantee the owner's copyright.

Successful segmentation of the text content is essential to this algorithm. Among all the languages, the segmentation of the text written in Chinese is the most difficult, because it has no apparent boundary to words in a text and the same word applied in different environments has different meanings. Our experiments used Chinese text. In this paper, we chose the most mature segmentation software-ICTCLAS-as the segmental tool to segment the text content and label the word's part of speech. Then assemble these segmented words into a feature string in order that the word appeared in the text. However, the dimension of text features will be extremely large if all segmented words

are assembled. So the reduction of dimensions is needed to original text features. We used the stop words database to remove all the words that have little value in expressing the theme of the text content. After that, we still removed some parts of the features according to the part of speech. It could improve the efficiency of the algorithm.

Although we have used the database of stop words and wiped out some of the valueless words, the dimension of the text features is still quite large. It is unfavorable for an excellent performance of the algorithm. So we further reduce the dimensions of features by the method of Latent Semantic Analysis. Latent Semantic Analysis (LSA) is a technique to compare texts using a vector-based representation that is learned from a corpus. LSA has both theoretical support and empirical results which show how it matches human behavior. The primary function of LSA is to compute the similarity of word pairs by comparing their vector representation. This relatively simple similarity metric has been situated within a psychological theory of text meaning and has been shown to closely match human capabilities on a variety of tasks.

The technology of Latent Semantic Analysis has several advantages: it can reduce the dimension of synonyms - words that have similar meaning; it also can wipe out the noise which is introduced in reducing the dimension process. Using this method would make the algorithm have much stronger robustness; it can also be applied to many kinds of languages.

After the process of Latent Semantic Analysis, which could reduce the dimensionality and improve the computing speed, we calculate the similarity value between the dimension-reduced text features which will be registered in the IPR database.

In this algorithm, we use the information the user provided and the secret key to ensure the safety of the copyright. In addition, the watermark construction with current time stamp can prevent the text file from being reconfigured by the attack later. In the last, the text features, the secret key, the user provided information and the current time stamp will be stored in the IPR database as a certification for the copyright. The structure of text copyright protection based on the text features shown as Figure2.

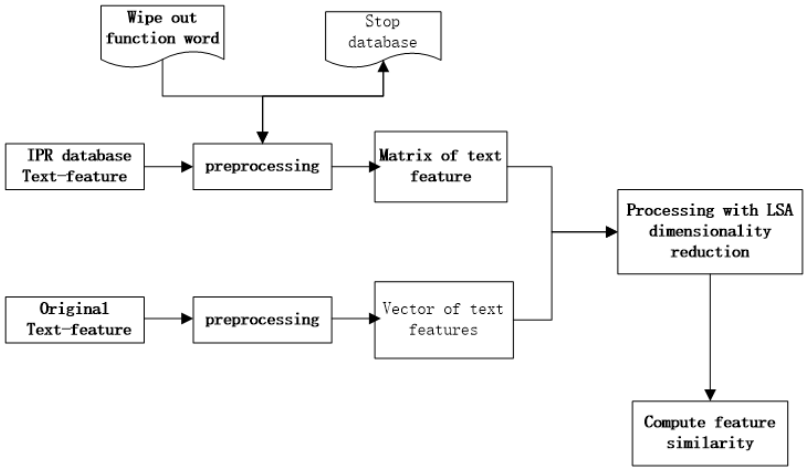


Figure 2. Structure of Text Copyright Protection Based on the Text Feature

3.1 Compute of Similarity

First, we define some symbols for the presentation of the process of calculating the similarity of two documents. The symbol of T stands for the text document, and the symbol of W_i is the feature item extracted from the text document T according to the word segmentation algorithm. The symbol of $F(W_i)$ represents the occurrence frequency

of the feature item W_i in the document *T*. The symbol of S_W is the sequence of all feature items assembled according to the order they appear in the document.

After getting the sequence of feature items and the information of the occurrence frequency, we can deduce the algorithm of calculating the similarity between two text documents, which can determine the degree of similarity of the two documents.

The symbol of T_1 represents the content of the text document to be detected, and T_2 is the content of the text document from the IPR to be compared withT_1. The symbol of $S_1(W)$ is the sequence of feature items of the text document to be detected. The number of items in $S_1(W)$ is K_1, and the symbols of $S_2(W)$, K_2 are corresponding to the text document of T_2. The inter section that contains the feature items belonging to both $S_1(W)$ and $S_2(W)$ is $S_1(W) \bigcap S_2(W)$, and the number of items in this set is K_W. The symbol of $Z(W_i)$ stands for anyone of feature items in the set of $S_1(W) \bigcap S_2(W)$, and the symbols of $F_1(W_i)$, $F_2(W_i)$ present the frequency of occurrence of the feature item in $Z(W_i)$ respectively in T_1, T_2. We use ρ1 stands for the degree of similarity betweenT_1 and T_2.In the light of the weakness of expressing the similarity by ρ1 alone, we can introduce ρ2 to improve the accuracy of the similarity of two text documents. The symbol ofρ2 is the extent of similarity of the occurrence frequency of feature items in the set of $S_1(W) \bigcap S_2(W)$.

As the stability of the consequent of MAX(S_1(W), S_2(W)) is not as excellent as the express of $[S_1(W) + S_2(W)]/2$, we define ρ1 as the equation (4):

$$\rho 1 = \frac{(K_1 \bigcap K_2)}{\frac{S_1(W) + S_2(W)}{2}} \tag{4}$$

This formula reflects what proportion the feature items in $S_1(W) \bigcap S_2(W)$ accounts for in the whole sequence of feature items. The extent of similarity of the occurrence frequency of each feature item in $S_1(W) \bigcap S_2(W)$ is reflected by the value ofρ2.

$$\rho 2 = 1 - \frac{\sum_{i=1}^{K_w} abs[F_1(W_i) - F_2(W_i)]}{\sum_{i=1}^{K_w} F_1(W_i) + \sum_{i=1}^{K_w} F_2(W_i)} \tag{5}$$

For the pair of values of ρ1, ρ2, we give weights of w1,w2 respectively according to their importance, thus, it generates the similarity of two text documents we want:

$$sim = \rho 1 * w1 + \rho 2 * w2 \tag{6}$$

3.2 Extraction and Recognition of Text Feature

In order to reduce the dimensions of the text features, except for the segmented contents of the text used in the database of stop words, it is necessary to select part of the segmented words according to the part of speech, such as nouns, verbs, adjectives, and to remove the valueless functional words. Then, using the method of Latent Semantic Analysis (LSA) to reduce dimensions of text features further. Finally, the remaining

segmented words and their frequencies will be assembled as text features. According to the needs of the works, we could also import some inherent words to the user dictionary before the text has been segmented; it contributes to extract text features which reflect the theme of the text.

3.2.1 Extraction of the Text Feature

Traditional watermarking algorithms are constructed with certain meaning on the basis of the text content. This proposed algorithm allows users to provide the information meaningless to watermarking, and stores the user's information and the secret key to guarantee the security of the information saved in the IPR database.

The algorithm extracts features according to the text content, regardless of what the format of the text is, such as Text Based PDF, Word (DOC & DOCX) and so on. It could also be used on the Plain Text (txt) and other documents that have no formats.

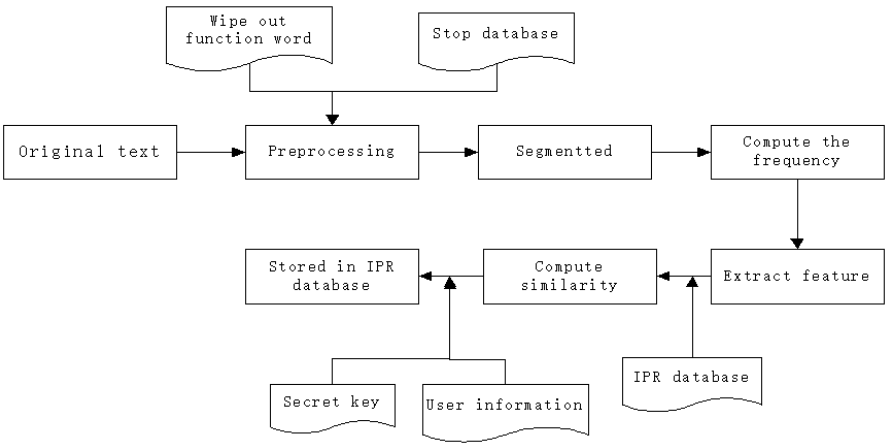


Figure 3. The Process of Extracting the Text Feature

The steps of this algorithm are as follows:

1) Obtain the content of the source document to be protected. After the preprocessing, word segmentation will be followed. Assemble all the segmented words into an orderly string according to the natural order of the text.
2) Select the text features of the document. Use the database of stop words to remove words which are valueless, and then, wipe out the grammatically-partial words which contribute less to express the theme of the text. Then, reduce the dimension of text features and compute each word's occurrence frequency.
3) In order to avoid reconstruction attack, the current time stamp will be saved in the IPR database in the form of string. For instance, the current time is June 18, 2012 22:18; it could be presented like 201206182218.
4) Similarity of the theme: Count the number of feature items in the public set $S_1(W) \cap S_2(W)$, which is the inter section of two text features from two text documents. Then, calculate the average length of $S_1(W)$ and $S_2(W)$ and get the similarity value $\rho 1$.
5) The similarity of the occurrence frequency of feature items: calculate the frequency of occurrence of feature items in the public set $S_1(W) \cap S_2(W)$, which is the inter section of two text features from two text documents T_1 and T_2, and get the similarity value $\rho 2$.
6) Assign weights to this two values: $sim = \rho 1 * w1 + \rho 2 * w2$.

7) If the similarity is less than the threshold value, it is indicated that the paper user provided has not been registered in IPR databases and it can be registered.
8) Users can construct the secret key based on the basic information. The front two bits are given by the system, and the remaining 5-8 bits are given by users.

The process of extracting the text feature is shown in Figure.3.

3.2.2 Recognition of the Text Feature

The proposed algorithm can extract copyright information even though the text suffered from attacks. The process of recognition is as follows.

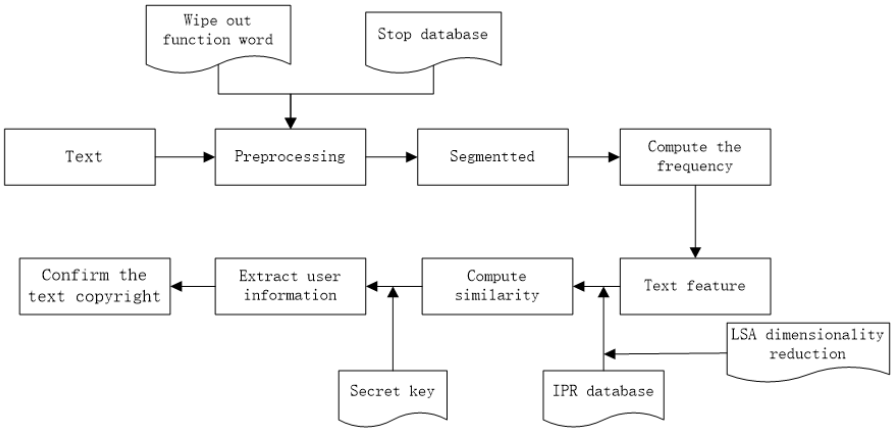


Figure 4. The Process of the Recognition of Text Features

1) Obtain the content of the source document to be verified, after the preprocessing, word segmentation will be followed. Assemble the segmented words into an orderly string according to the natural order of the text.
2) Select the text features of the document. Use the database of stop words to remove words which are valueless, and then wipe out the grammatically-partial words which contribute less to express the theme of the text. Then, reduce the dimension and compute each word's occurrence frequency.
3) Compute the similarity value $\rho 1$ and $\rho 2$ between text features of the source document the user provided and features of the text stored in the IPR database. Assign two weights to the similarity value $\rho 1$ and $\rho 2$.
4) Compare the similarity with the threshold value. If it is larger than the threshold, it is indicated that the document user provided has been registered in the IPR database. If the user can provide the right information such as the secret key and author name, then the system will display the information the user stored when the text registered to verify the ownership of copyright. Otherwise, the information the user provided is not completed to verify the copyright. If the similarity is smaller than the threshold, the text can be registered to make copyright be protected.

The process of recognition of the text feature is shown as Figure4.

4. Experimental Results and the Analysis of the Performance

4.1 Ascertain the Threshold

200 text samples were selected randomly. The similarity value between the original text and the test sample was computed respectively by using the proposed algorithm in this paper. The experiment result is shown in Figure 5. This experiment indicates that the

similarity values are almost distributed between 0.1 and 0.4. In order to reduce the probability of false detection, in this paper, the threshold is enlarged to 0.7.

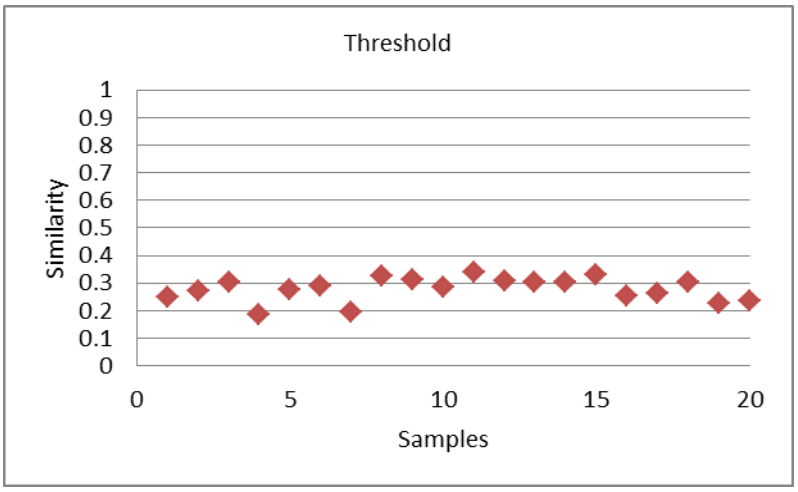


Figure 5. Ascertain the Threshold

4.2 Attack Experiments

This system is developed on the platform of Eclipse with java language, and it can be applied to many kinds of text formats. For instance, documents we have seen the most: Text Based PDF, Word (DOC &DOCX), Plain Text and so on, can be experimented in this system. In order to test the robustness and accuracy of the algorithm, we make attack experiments. For example, adjust the format, modify the key and information, delete part information, adjust the location of the paragraph *etc.*

The experiment of text features extraction and recognition was executed (or carried out on) on doc format. This system includes two parts: document registering and copyright verification. The original document must pass through the examination whether it has a copyright or not if it needs to be registered. Namely, the original document features should be compared with documents features in the database of IPR. When the similarity value is less than the threshold, this document will be registered. The part of copyright verification is provided for users. The document registered in the IPR can be extracted out of the watermark information with the accurate information provided: author name and key.

The experiments of the attack to the text format include changing the format of the document, attacking text with font adjustment, deleting paragraphs and spaces. The result of the attack experiments is as shown in the Table1.We can find that it can extract watermark completely. It indicates that this algorithm has a strong resistance to the format attack. When comparing this algorithm with the similar algorithm proposed by Qin [18] in format attack, with the increase of the adjusting scale, the robustness of Qin's algorithm [18] has decreased sharply. However, the algorithm in this paper has a stable performance in this experiment. The experiment result is shown on Figure6.

Table 1. Attack of Format

Attack	Similarity
Adjust text font and paragraph	1
Transform the text format	1
Delete Spaces	1

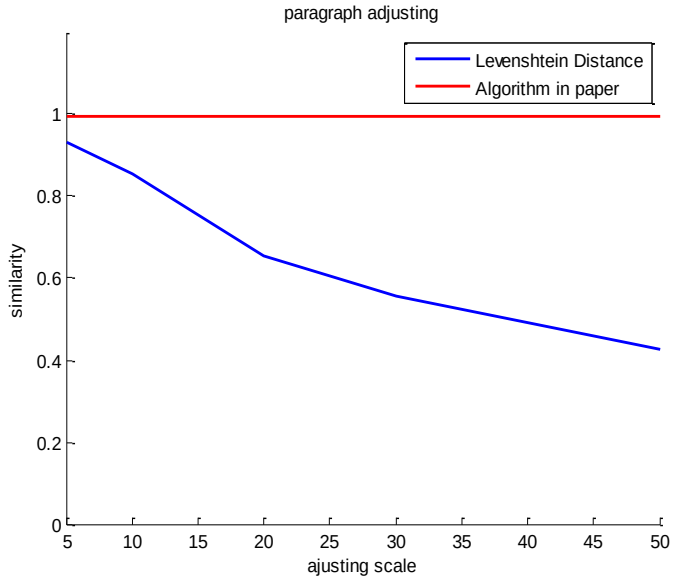


Figure6. Compare the Format Attack between Our Algorithm and Ref. [9]

Attacks to the text content include paragraph deleted, content added, and part of the content changed. Part of the content changed includes: modification of key words of the content and synonym replacement and so on. Experiments were done to 60 papers with different length, and every kind of attack experiment has been done 20 times, here takes the mean value as the last result. The experiment result is shown in the Table2. We compare the attack results with those of similar algorithms [17-18], the experiment result is shown on Figure7. It indicates the algorithm in this paper has stronger robustness than the similar algorithms.

Table2. Attack to Content

Attack styles	percentage	Similarity
Part of the content deleted	5	98.5
	10	97.1
	20	93.3
	30	89.7
	50	83.2
Content added	5	97.9
	10	96.1
	20	93.3
	30	90.7
	50	85.1
Part of the content changed	5	96.8
	10	93.3
	20	87.8
	30	83
	50	73.1

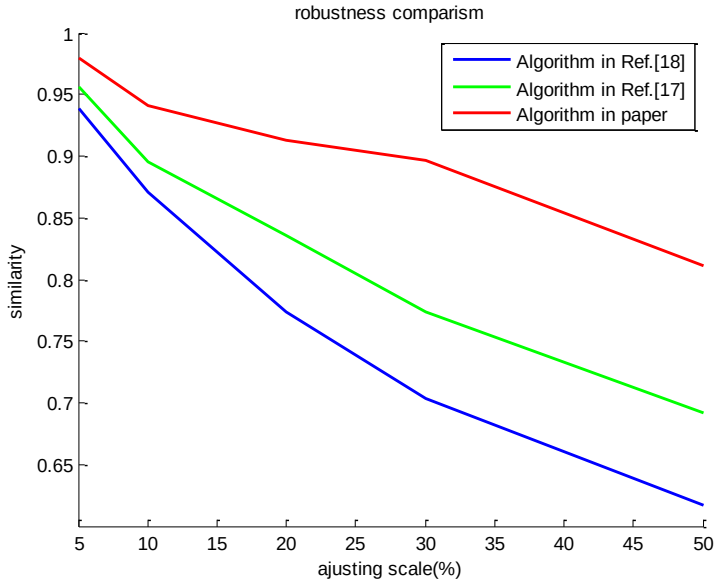


Figure 7. The Robustness Compared with Similar Algorithms

The experiment results show better invisibility comparing with the other text-based algorithm. In the aspect of transparency, this algorithm has an obvious superiority compared with the watermarking algorithm based on natural language. It has a more brilliant performance in obtaining the theme features than the traditional algorithm that based on the structure of Chinese character. As to similar algorithms in Ref. [17-18], the algorithm proposed in this paper has stronger robustness.

1) Transparency**:** It is different from the traditional text watermarking algorithm. Text feature-based algorithm is transparent completely; it means this algorithm will do nothing modified to the text content.
2) Robustness: The algorithm must have a certain ability to resist attacks: such as simply rearranging the content of the text, converting the format of the text, modifying character features, which can be able to break the feature of the text. This algorithm proceeds from the frequency of feature items that can represent the theme of the text. It extracts features on the basis of the frequency and the part of speech of feature items, so it avoids the influence from the format of the document, the feature of characters, and the number of blank. It also has a strong robustness to attacks like: deletion and addition of parts of the text, and the malicious tampering of the text.
3) Capacity**:** Little redundancy is the biggest drawback of the text, so it leads to a small capacity of text watermarking. The algorithm in this paper solved this problem, we use the author name, title, and some other information the user provided together as the watermark. Moreover, this algorithm has used the time stamp mechanism which can make it resist text reconstruction attack well.

The algorithm also has very strong robustness to the text contents image. No matter what kinds of attacks that have been done to the image in the text documents, the algorithm in this paper still can confirm the text copyright effectively.

5. Conclusions

Due to little researches on text watermarking, this paper has proposed a new text watermarking algorithm to calculate the similarity. In this paper, the text features were constructed by segmented words and the occurrence frequency of each word. There are two steps to realize the dimension reduction. The first step is to import the database of

stop words and wipe out the words according to the part-of-speech, and to select the words that contribute most to the theme of the text. And the second step uses the technology of LSA to reduce dimensions. It has not only achieved the target of reducing dimension in entirety, but also has improved the efficiency of the system. The speed of processing the text has also improved. After the process of dimension reducing, the system will combine the user provided information with the text feature to store in the IPR database.

The algorithm has advantages compared with previous watermarking algorithms. The experiment has proved that the method has a stronger robustness and better transparency, and it also has more embedding capacity. It has solved the contradiction between watermarking capacity and robustness. The algorithm could be applied to many kinds of languages, and the probability of collision of the watermark carrier with different features is almost zero.

Acknowledgements

This research project was supported by the National Natural Science Foundation of China (Grant No. 61262090)

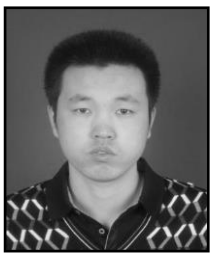
References

[1] Y. Z. Cheng and J. Zhang, "An algorithm for the illegal copying detection of digital documents", Proceedings of the IEEE International Conference on NLP-KE, Wuhan, China, **(2005)**, pp. 384-387.
[2] Z. Dawei, C. Guanrong and L. Wenbo, "A chaos-base robust wavelet-domain water marking", Chaos, Solutions & Fractals, vol. 22, **(2004)**, pp. 47-54.
[3] T. M. Ng and H. K. Carg, "Maximum-likelihood detection in DWT domain image watermarking using Laplace modeling", Signal Processing Letters, IEEE, vol. 12, no. 4, **(2005)**, pp. 285-288.
[4] S. Agreste, G. Andaloro, D. Prestipino and L. Puceio, "An image adaptive, wavelet-based watermarking of digital image", Journal of Computational and Applied Mathematics, In Press, Corrected Proof, Available online, **(2007)**, pp. 13-21.
[5] H. Cheng, "A review of video registration methods for watermark detection in digital cinema applications", in: 2004 IEEE International Symposium on Circuits and Systems, Vancouver, BC, Canada, vol. 5, **(2004)**, pp. 704-707.
[6] M. Pooyan and A. Delforouzi, "Adaptive and robust audio watermarking in wavelet domain", Intelligent Information Hiding and Multimedia Signal Processing, Third International Conference, **(2007)**, pp. 10-13.
[7] Shang qinxiao. The study of robust digital watermarking.HaZhon UniversityDoctor degreepaper.(2008), pp. 1-10.
[8] M. J. Atallah, V. Raskin and M. Crogan, "Natural language watermarking: Design, analysis, and a proof-of-concept implementation", Proceedings of the 4th International Workshop on Information Hiding. Springer-Verlag, London, **(2001)**, pp. 185-199.
[9] M. J. Atallah, C. J. Mcdonough and V. Raskin, "Natural language proceeding for information assurance and security: An overview and implementations", Proceedings of the 2000 Workshop on New Security Paradigms. ACM Press, New York, **(2000)**, pp. 51-65.
[10] J. Brassil, S. Low and F. Maxemchukn, "Copyright protection for the electronic distribution of text documents", Proceedings of the IEEE, vol. 7, no. 89, **(1999)**, pp. 1181-1196.
[11] J. Brassil, S. Low and F. Maxemchukn, "Electronic marking and identification techniques to discourage document copying", IEEE Journal on Selected Areas in Communications, vol. 8, no. 13, **(1995)**, pp. 1495-1504.
[12] Y. Z. Cheng and J. Zhang, "An algorithm for the illegal copying detection of digital documents", Wuhan, China: Proceeding of the IEEE International Conference on NLP-KE, **(2005)**, pp. 384-387.
[13] X. M. Sun, G. Luo and H. J. Huang, "Component-based digital watermarking of Chinese texts", Shanghai: Proceeding of the Third International Conference on Information Security, **(2004)**, pp. 76-81.
[14] I. B. Ozer, M. Ramkumar and A. N. Akansu, "New method for detection of watermarks ingeometrically distorted images", 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'), Istanbul, Turkey, **(2000)**, pp. 1963-1966.
[15] C. Jang and X. Chen, "A robust text algorithm based text content", The paper on automatic, vol. 36, no. 9, **(2012)**, pp. 1250-1254.
[16] X. Wang and C. Yan, "Text zero-watermarking based the Chinese frequency", Computer apply, vol. 9, no. 26, **(2009)**, pp. 2366-2341.
[17] J. Shu and Y. Liu, "Text zero-watermarking algorithm based Chinese word frequency", Computer apply,

vol. 2, no. 31, **(2011)**, pp. 104-105.
[18] Q. Si, L. Zhang and D. Lian, "Text watermarking algorithm base feature text", Computer apply, vol. 9, no. 29, **(2009)**, pp. 2348-2350.
[19] J. Zunera, M. M. Anwar and J. Hajira, "Word length based zero-watermarking algorithm for tamper detection in text documents", 2010 2nd International Conference on Computer Engineering and Technology. Washington, DC: IEEE Computer Society, vol. 6, **(2010)**, pp. 378-382.
[20] J. Zunera, M. M. Anwar and S. Maria, "Content based zero-watermarking algorithm for authentication of text documents", International Journal of Computer Science and Information Security, vol. 2, no. 7, **(2010)**, pp. 212-217.
[21] L. Pan and J. Zou, "Text zero-watermarking based English text content", Conference communication of the 12th national youth. Publishing of electronics industry, Beijing China, **(2007)**.
[22] Y. Cheng, X. Sun and H. Huang, "A zero-watermarking algorithm based chaos mapping", electronic paper, vol. 2, no. 31, **(2003)**, pp. 214-216.
[23] H. Ding and Y. Hong, "Inter word distance changes represented by sine waves for watermarking text", IEEE Trans on Circuits and Systems for Video Technology, vol. 12, no. 11, **(2001)**, pp. 1237-1245.
[24] Y. Meng, C. Wu and S. Su, "A discuss for database with zero-watermarking", Computer Science, vol. 32, **(2005)**, pp. 381-383.

Authors

De Li, received the Ph.D. degree from Sangmyung University, major in computer science in 2005. He is currently a professor of Dept. of Computer Science at Yanbian University in China. He is also a Principal Researcher at Copyright Protection Research Institute, Sangmyung University. His research interests are in the areas of copyright protection technology, feature recognition, digital watermarking, and digital forensic marking.



Xue Zhe Jin, is a postgraduate, major in Information Security, now studying at Yanbian University in China. His research interests are in the areas of copyright protection technology, information security, zero watermarking.

Li Hua Cui, received the Ph.D. degree from KookMin University, major in Financial Management in 2008. She is currently a professor of Dept. of Financial Management at Yanbian University in China. Her research interests are in the areas of Statistical analysis, Information hiding, Pattern recognition, copyright protection technology.

