Soft Comput (2006) 10: 41–46
DOI 10.1007/s00500-005-0463-2

FOCUS

Jun-Peng Bao · Jun-Yi Shen · Hai-Yan Liu
Xiao-Dong Liu

A fast document copy detection model

Published online: 6 April 2005

Abstract Text similarity measure is a common issue in Information Retrieval, Text Mining, Web Mining, Text Classification/Clustering and Document Copy Detection etc. The most popular approach is word frequency based scheme, which uses a word frequency vector to represent a document. Cosine function, dot product and proportion function are regular similarity measures of vector. But they are symmetric similarity measures, which cannot find out the subset copies. In this paper we present the concepts of asymmetric similarity model and heavy frequency vector (HFV). The former can detect subset copies well, and the latter can save a great resources and CPU time. We have developed two new asymmetric measures: heavy frequency vector (HFM) and Heavy inclusion proportion model HIPM. The HFM and HIPM are derived from cosine function and proportion function by combining asymmetric similarity concept with HFV. The HFV is to truncate the original full frequency vector to a short vector. We can adjust the parameter of HFV to balance the model's performance. The paper illustrates the aspects of asymmetric similarity and HFV models by several experiments.

1 Introduction

According to the similarity of documents, the relations between two text documents are: (1) unrelated, they are very different, nothing shared, (2) related, they are a little similar but it is clear that they are different documents. For example, two documents discuss the similar topics in very different words, (3) partly overlapped, some sections of them are (almost) the same while the rest are different, (4) subset, one document is a part of the other, (5) copy, they are the same. It is obvious that the most similar documents are the same documents and the unrelated documents are not similar at all.

Documents similarity (or relationship) measure is a common and elementary issue in Information Retrieval (IR), Text Mining (TM), Web Mining (WM), Text Classification (TC)/Clustering and Document Copy Detection (DCD). In IR we need to find out similar papers according to the given sample. In TC variant similarity measures are key factor to affect performance. In these domains we identify text in a loose boundary because we need recall as many as related documents. But in DCD we do it more accurately so as to distinguish copies from a lot of similar documents.

In this paper we present the concepts of asymmetric similarity model and heavy frequency vector (HFV) in order to detect plagiarized documents in corpus. The former can detect subset copies well; the latter can save a great resources and CPU time. Section 2 introduces some related work. In sect. 3 we discuss similarity measure models in detail. In sect. 4 we introduce contrast experiments of several models. In sect. 5 we discuss features of different models. At last we draw conclusions in sect. 6.

2 Related Work

Brin et al. (1995) built the first document copy detection system: copy protection system (COPS), which detects document overlap based on sentence and string matching but it cannot find partial sentence copy. Before long, Garcia-Molina and Shivakumar and Garcia-Molina (1995) developed stanford copy analysis method (SCAM) to improve COPS. In SCAM they present relative frequency model (RFM) so as to find out subset copies. The RFM is the first asymmetric similarity measures in copy detection. Later they presented other prototypes based on SCAM to extend detection range

J.-P. Bao (✉) · J.-Y. Shen · H.-Y. Liu · X.-D. Liu
Department of Computer Science and Engineering,
Xi'an Jiaotong University, Xi'an,
710049, People's Republic of China
E-mail: baojp@mail.xjtu.edu.cn

from a single register database to distributed databases and the WEB (Garcia-Molina et al. 1996; Sivakumar and Garcia-Molina 1995).

Heintze (1996), Broder et al. (1997), and Monostori et al. (2000) proposed methods similar to COPS. They made various approaches to find strings as "fingerprints", and detect plagiarism according to the common fingerprints proportion.

Si (1997) built a copy detection mechanism called CHECK, which was some like SCAM. They both adopted many IR techniques and detected overlap based on word frequency. But CHECK parsed each document to build an internal indexing structure called structural characteristic (SC) and they detected plagiarism according to the key words proportion of SC nodes.

Song and Shen (2001) proposed an illegal copy and distribution detection algorithm for digital goods (CDSDG). Their algorithm indeed combined the SC structure of CHECK and the RFM of SCAM to discovery illegal copy content.

3 Similarity measure models

3.1 Symmetric and asymmetric similarity measures

There are two major purposes of text similarity measure: the one is find out all of the similar (or related) documents from a large document collection, such as IR, TM, WM and TC; the other is to find out the copies (plagiarisms) of a document from the collection, such as DCD. That causes different requests for similarity measure. The former wants to return those related documents that are far away from other categories. The latter distinguishes almost the same documents (copies or plagiarisms) from other similar documents.

Cosine function, dot product and proportion function are commonly used similarity measures. Usually, we define the similarity value in [0,1] so that cosine and proportion functions are used more. Let $F(A)$ and $F(B)$ be document A and B word frequency vectors, then the similarity of A and B in cosine function is $S_{\cos}(A,B)$:

$$S_{\cos}(A,B) = \frac{\sum_{i=1}^{n}\left[\alpha_i^2 \times F_i(A) \times F_i(B)\right]}{\sqrt{\sum_{i=1}^{n}\left[\alpha_i^2 \times F_i^2(A)\right] \times \sum_{i=1}^{n}\left[\alpha_i^2 \times F_i^2(B)\right]}} \tag{1}$$

where α_i is the word weight vector, $F_i(A)$, $F_i(B)$ are the respective number of occurrences of the ith word in A and B. It is obvious that $S_{\cos}(A,B) = S_{\cos}(B,A)$. The similarity of A and B in proportion function is $S_{\%}(A,B)$:

$$S_{\%}(A,B) = \frac{|F(A) \cap F(B)|}{|F(A) \cup F(B)|} = \frac{\sum_{i,j=1}^{n} \alpha_i (F_i(A) \oplus F_j(B))}{\sum_{i=1}^{n} \alpha_i F_i(A) \sum_{j=1}^{n} \alpha_i F_i(B)} \tag{2}$$

where $F_i(A) \oplus F_j(B)$ means that:

$$F_i(A) \oplus F_j(B) = \begin{cases} F_i(A) + F_j(B) & w_i = w_j \\ 0 & w_i \neq w_j \end{cases} \quad i,j = 1,2,3,\cdots,n \tag{3}$$

It is obvious that $S_{\%}(A,B) = S_{\%}(B,A)$.

Because the similarity of A to B and that of B to A is the same in above two measures, i.e. $S(A,B) = S(B,A)$, we call them symmetric similarity. For symmetric similarity, the copies' (same documents) value is 1 and the more overlapped words between documents the higher score. But they cannot distinguish the subset copies from partly overlapped documents. As we know that A is included in B is different from B is included in A, i.e. $A \subset B \neq B \subset A$. So the inclusion measure of $A \subset B$ should be different from that of $B \subset A$. However the symmetric similarity does not satisfy that.

In order to find out subset document copy, Garcia-Molina et al. (1996) and Shivakumar and Garcia-Molina (1995) proposed relative frequency model (RFM). The subset measure of document A to be a subset of document B to be:

$$\text{Subset}(A,B) = \frac{\sum_{w_i \in c(A,B)} \alpha_i^2 \times F_i(A) \times F_i(B)}{\sum_{i=1}^{N} \alpha_i^2 F_i^2(A)} \tag{4}$$

It is obvious that $\text{Subset}(A,B) \neq \text{Subset}(B,A)$ and $\text{Subset}(A,A^c) = 1$ if A^c is a copy of A. Hence we call this type as asymmetric similarity measure. The RFM similarity measure between two documents A and B is:

$$S_{\text{RFM}}(A,B) = \max\{\text{Subset}(A,B), \text{Subset}(B,A)\} \tag{5}$$

The $\text{Subset}(A,B)$ may be greater than 1. In order to regularize the similarity value in [0,1], the final RFM similarity of documents A and B is:

$$S_{\text{RFM}}(A,B) = \min\{1, \max\{\text{Subset}(A,B), \text{Subset}(B,A)\}\} \tag{6}$$

The RFM is derived from cosine function. Similar to that, we define another asymmetric similarity that derived from proportion function. We call it inclusion proportion model (IPM) . The inclusion proportion of $A \subset B$ is:

$$\text{Incl}(A,B) = \frac{|F(A) \cap F(B)|}{|F(A)|} = \frac{\sum_{i,j=1}^{n} \alpha_i (F_i(A) \oplus F_j(B))}{2 \times \sum_{i=1}^{n} \alpha_i F_i(A)} \tag{7}$$

Obviously, $\text{Incl}(A,B) \neq \text{Incl}(B,A)$ and $\text{Incl}(A,A^c) = 1$ if A^c is a copy of A. Similar to RFM, the final IPM similarity of documents A and B is:

$$S_{\text{IPM}}(A,B) = \min\{1, \max\{\text{Incl}(A,B), \text{Incl}(B,A)\}\} \tag{8}$$

According to experiments, we believe the RFM and IPM are both excellent for subset copy detection.

3.2 Text vectors

All of the above measures are based on word frequency vector of a text document, and only the frequencies of the same words in two documents can be operated. These representation methods are derived from vector space model (VSM). In the original VSM, each word's frequency is contained in the vector so that a document word frequency vector is a very huge vector, which dimensions are often thousands. Assuming a document in a corpus contains k different words in average and there are n documents altogether in the corpus. Then we must compare k times at least on the whole document to find out all the same words between two documents, and $k(n\text{-}1)$ times at least in the total corpus in order to get the similarities of the document with all others. We call this kind of text vector as full frequency vector (FFV). Therefore it is important to reduce the huge FFV so as to save resources, increase speed and improve performance.

The RFM vector is a FFV, but not all frequencies of the vector contribute to the RFM similarity. Garcia-Molina et al (1996); Shivakumar and Garcia-Molina (1995) define a closeness set $c(A,B)$ to contain those words w_i that have similar number of occurrences in the two documents A and B. That is, a word w_i is in $c(A,B)$ if it satisfies the following condition:

$$\varepsilon - \left(\frac{F_i(A)}{F_i(B)} + \frac{F_i(B)}{F_i(A)}\right) > 0 \tag{9}$$

where $\epsilon = (2^+, \infty)$ is a user-tunable parameter. Only the words that is in $c(A,B)$ have effect on the RFM similarity of A and B. The value of $\sum_{i=1}^{N} \alpha_i^2 F_i^2(A)$ is fixed to document A, but the closeness set of A is different while B is different. That is to say we have to get the closeness set for each document pair. Consequently, the complexity of RFM in the whole corpus is $O(k(n-1))$ because the RFM vector is a FFV.

The main purpose of RFM is to find out copies and subset copies from a lot of similar documents. There are two facts in plagiarism documents:

1. If there are plenty of the same words between two documents, then they may be plagiarism or sharing almost the same topics at least.
2. The more similar distribution (proportion) of the common words, the more documents similarity.

The cosine function satisfies (1) but not (2), whereas RFM measure satisfies both. So RFM is better than cosine, and it is successfully used to track real life instances of plagiarism in several conference papers and journals (Denning 1995). But FFV is low efficient, we present the heavy frequency vector (HFV) to improve performance.

3.3 Heavy frequency vector

As well known, we can omit the tiny parts of a function to get a proximate result and gain great performance improvement. For instance, e equals to:

$$e = 1 + 1/1! + 1/2! + \cdots + 1/n! + \cdots \tag{10}$$

In fact, we often omit the tiny parts to get a proximate result within certain precision, e.g.

$$e = 1 + 1/1! + 1/2! + \cdots + 1/10! \tag{11}$$

This is a practical approach. We believe that the small frequency words are tiny parts of a document, so we delete them from the full text vector. Thus we could get a fixed length word frequency vector of a document, which contains only the high frequency words and no low frequency words. We call it as HFV. The dimension of a HFV is α parameter, which is a threshold to HFV. Namely, we take the top α frequent words in a document to construct the HFV. It is obvious that the HFV of a document is shorter than FFV. That is the real reason that we get great improvement. Assuming a corpus that contains n documents and each document has k different words in average, then the complexity of HFV in the whole corpus is $O(\alpha\ (n-1))$. Because $\alpha \ll k$, we can save more resources and running time.

Applying the HFV concept to asymmetric similarity measure, we get two new asymmetric similarity measure models: heavy frequency model (HFM) and heavy inclusion proportion model (HIPM). These two models have almost the same performance in text identification. In HFM the subset measure of $A \subset B$ is:

$$\mathrm{Subset}_{\mathrm{HFM}}(A,B) = \frac{\sum_{i,j=1}^{n} \{[F_i(A) \otimes F_j(B)] \times [1 - |P_i(A) - P_j(B)|]\}}{\sum_{i=1}^{n} F_i^2(A)} \tag{12}$$

The final HFM similarity is:

$$S_{\mathrm{HFM}}(A,B) = \min\{1, \max\{\mathrm{Subset}_{\mathrm{HFM}}(A,B), \mathrm{Subset}_{\mathrm{HFM}}(B,A)\}\} \tag{13}$$

where we select the most frequent n (e.g. 100) words in document D as its words frequency vector $F(D)$, i.e. $F(D)$ is a HFV. $P(D)$ is the respective proportion vector and $P(D) = F(D)/|D|$, namely, is a HFV too. $|D|$ is the amount of words in D. The operator $\otimes$ is defined as follows:

$$F_i(A) \otimes F_j(B) = \begin{cases} F_i(A) \times F_j(B) & w_i = w_j \\ 0 & w_i \neq w_j \end{cases} \quad i,j = 1,2,3,\cdots,n \tag{14}$$

In HIPM the inclusion proportion of $A \subset B$ is:

$$\mathrm{Incl}_{\mathrm{HIPM}}(A,B) = \frac{\sum_{i,j=1}^{n} \{[F_i(A) \oplus F_j(B)] \times [1 - |P_i(A) - P_j(B)|]\}}{2 \times \sum_{i=1}^{n} F_i(A)} \tag{15}$$

The final HIPM similarity is:

$$S_{\mathrm{HIPM}}(A,B)=\min\{1,\max\{\mathrm{Incl}_{\mathrm{HIPM}}(A,B),\mathrm{Incl}_{\mathrm{HIPM}}(B,A)\}\} \tag{16}$$

where the operator $\oplus$ is the same with IPM and $S_{\%}(A,B)$, the other notations are the same with HFM.

It is noticed that in HFM and HIPM we use the difference of word proportion as the word frequency vector's weight, namely the smaller difference of word proportion the higher weight and the larger difference the lower weight. Hence both of HFM and HIPM satisfy the two facts of plagiarism so that we can expect them to perform well in copy detection.

4 Experimental Results

4.1 Experiment Preparations

In this section we perform experiments to test our new models. We use the proceedings of international conference on machine learning and cybernetics 2002 (ICMLC2002) as test corpus, from which we select 483 papers that is longer than 4k. The original PDF documents are converted into plain ASCII text files, and then we use these text files to create four types copies: (1) Exact self-copy, these copies are just renamed source files. This type is denoted by *D-copy*. (2) Self cross copy, first we divide each source file into ten blocks and then we randomly reassemble these blocks into a new copy. This type is denoted by *O-copy*. (3) Subset copy, we select two files from the source files and divide each of them into ten blocks, at last we randomly reassemble these 20 blocks into a new copy. This type is denoted by *Y-copy*. (4) Partial copy, we select n (we set it to five in our experiment) files from the source files and divide each of them into ten blocks, then we randomly select k (it is two in our experiment) blocks from each ten blocks, at last we randomly reassemble all selected blocks into a new copy. This type is denoted by *X-copy*.

We make a document pair with a copy file and a source file and detect plagiarism between them. We define the positive error as the proportion of non-copy document pairs (no plagiarism) above the threshold τ in the whole non-copy document pairs, i.e.

$$E^{+}_{\Phi}(\tau)=\frac{|\{\Phi_N\}_{\geq\tau}|}{|\Phi_N|} \tag{17}$$

where Φ is some type of document pairs, and $\{\Phi_{\mathrm{N}}\}_{<\tau}$ is those non-copy pairs of type Φ whose plagiarism score is greater than or equal to τ. The negative error is the proportion of copy document pairs (containing plagiarism) below the threshold τ in the whole copy pairs, i.e.

$$E^{-}_{\Phi}(\tau)=\frac{|\{\Phi_C\}_{<\tau}|}{|\Phi_C|} \tag{18}$$

where $\{\Phi_C\}$is those copy pairs of type Φ whose score is less than τ.

4.2 Contrast experiments

4.2.1 Symmetric similarity versus asymmetric similarity

In this section we compare symmetric similarity measures with asymmetric similarity measures. In order to save time, we applying HFV to cosine and proportion functions to get two symmetric similarity measures of HFV: Hcos model and H% model. The Hcos similarity of document A and B is:

$$S_{\mathrm{Hcos}}(A,B)=\frac{\sum_{\mathrm{i,j}=1}^{n}\{[F_i(A)\otimes F_j(B)]\times[1-|P_i(A)-P_j(B)|]\}}{\sqrt{\sum_{\mathrm{i}=1}^{n}F_i^2(A)\times\sum_{\mathrm{j}=1}^{n}F_j^2(B)}} \tag{19}$$

where the notations are the same with HFM. The H% similarity of document A and it B is:

$$S_{\mathrm{H\%}}(A,B)=\frac{\sum_{\mathrm{i,j}=1}^{n}\{[F_i(A)\oplus F_j(B)]\times[1-|P_i(A)-P_j(B)|]\}}{\sum_{\mathrm{i}=1}^{n}F_i(A)+\sum_{\mathrm{j}=1}^{n}F_j(B)} \tag{20}$$

where the notations are the same with HIPM. Correspondingly, we select HFM and HIPM representing asymmetric similarity measures.

Figure 1 illustrates the four measures' positive error and negative error trend against the threshold on the

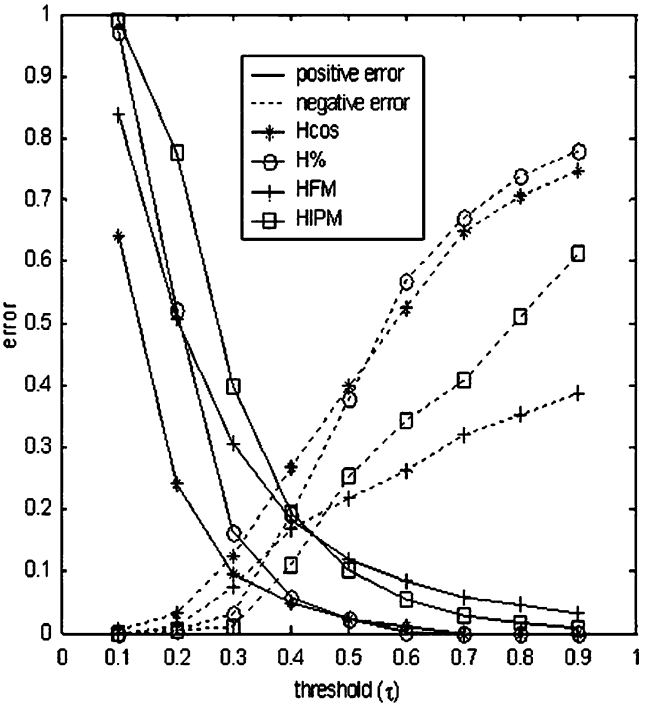

Fig. 1 Symmetric similarity measures vs. asymmetric similarity measures

whole corpus. Here we set the value of α parameter to 100. From Fig. 1, we know that in general positive errors of symmetric similarity are lower than that of asymmetric measures while negative errors of symmetric measures are higher than that of asymmetric measures. We find that the symmetric and asymmetric measures have almost the same error on the copy files (exact copies and self-cross copies) and the partial copy files, but asymmetric measures have much less negative error on the subset copy files at the expense of a little increase of positive error.

4.2.2 FFV versus HFV

In this section we compare FFV with HFV. We select cosine function, proportion function and RFM as delegations of FFV, Hcos, H% and HFM representing HFV. The Fig. 2 shows the error trends on the whole corpus of the six measures. The Fig. 3 illustrates all seven measures' running time in milliseconds against different corpus size, which are varied from 10 to 910. Its *Y*-axis is in log scale and the running time is the average of four tests.

From Fig. 2, we find that the proportion function has the sharpest error trends and RFM is the flattest. When the threshold is lower than 0.6, the RFM has the lowest positive error and the highest negative error, the proportion function has the highest positive error and lowest negative error. When the threshold is greater than 0.6, the HFM positive error is the highest, the others is declining to near 0; on the other hand the HFM negative error is the lowest. The Hcos and H% positive error is lower than cosine and proportion function positive error while their negative errors are in reverse situation. From figure 4.3, we see that the HFV model's running time is greatly lower than that of FFV.

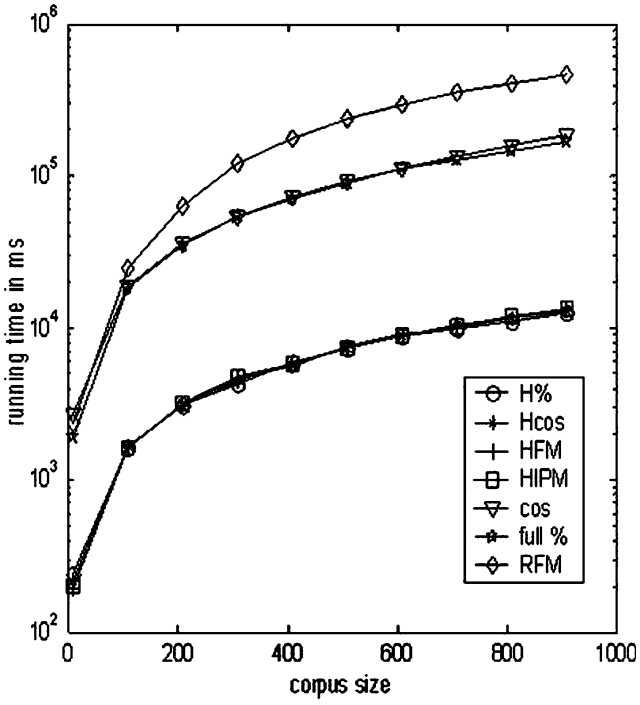

Fig. 3 The running time of various measures

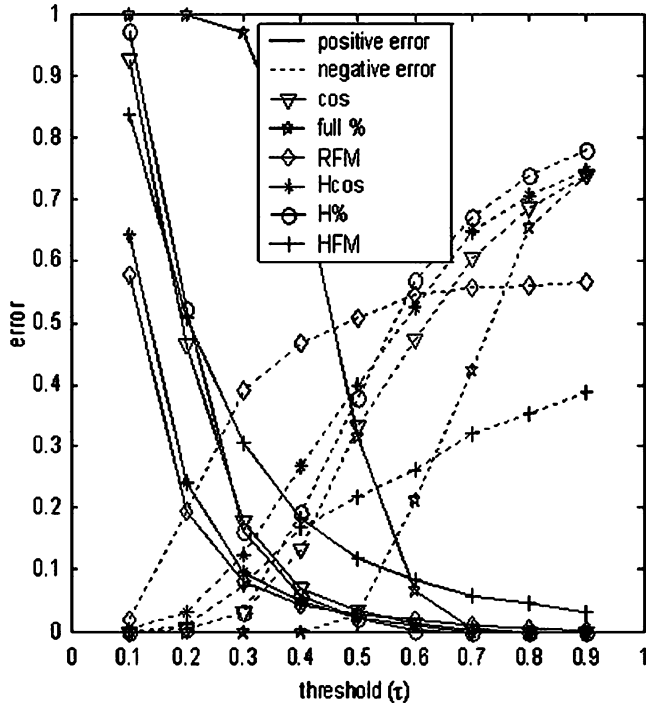

Fig. 2 FFV Measures vs. HFV measures

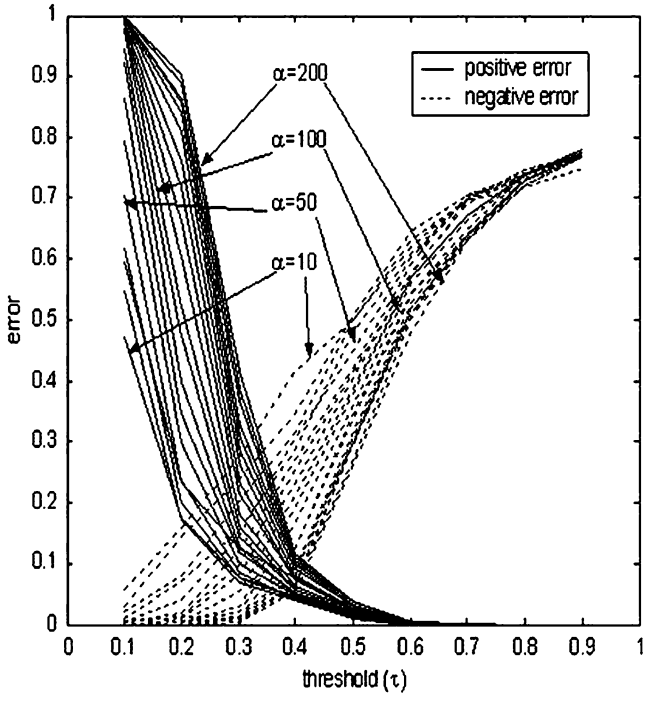

Fig. 4 The relationship between α parameter and errors

4.3 Effect of α parameter

We have mentioned in Sect. 3 that the α parameter defines the size of HFV so that it has effect on the last similarity score. In all of the above experiments, we set the α value to 100. The Fig. 4 illustrates the relationship between α parameter and H% error. The effect on other HFV models is the same as it. We vary the α value from 10 to 200 and find that as the α value growing the positive error is growing too, whereas the negative error is decreasing. There is another fact that the α value growing makes the similarity score growing so that the threshold has to increase to keep an error level. Because the high α value leads to high positive error, we cannot set the α value too large so that we need not make a large document vector. That means that we can detect fewer words without increasing error. It is very important for system to save index space and make a great improvement in efficiency. On the other hand, small α value makes negative error growing, that is to say we must test adequate words to ensure the validity of detection. Consequently, we suggest that it is better to set α value in [80,150].

5 Discussions

The positive error and the negative error are a pair of contradictory. None of above measures has both low positive error and negative error. That is to say a low positive error will lead to a high negative error and vice versa. Therefore we have to make a trade-off between the positive and negative error when we select similarity measures. If we are conservative, we can select low positive error measures such as RFM, Hcos, H% and cosine function. Otherwise we can select HFM, HIPM and proportion function.

The asymmetric models tend to make positive error to grow a little comparing with symmetric models, because asymmetric models make the similarity score growing. In contrast, the HFV models tend to decrease positive error comparing with FFV models. In fact, a shorter vector will lead to lower positive error and higher negative error. The key words can be considered as the extremely short vector (i.e. set α parameter to a extremely small value) and the FFV can be considered as the other extreme. The great virtues of HFV are that it saves much more resources and CPU time. Whatsoever we can balance the positive error, negative error, store space and CPU time by adjusting the α parameter.

All measures can distinguish copies (exact self copies and self cross copies) from similar documents, but only asymmetric models (RFM, HFM and HIPM) can find out subset copies. All these measures are based on word frequency, which has poor performance on partial copy. For partial copy, the common words are few in the whole vector so that the similarity score is low. Consequently, it cannot distinguish partial copies from related or similar documents. We believe that word frequency based approaches is not suitable for partial copy detection. The partial copies must have a few common sequences, therefore we can use some string matching approaches to find them.

6 Conclusions

Text identification is to identify the relationship between two text documents. To measure the similarity of two documents is a regular method to decide their relationship. The basic similarity measures of document vectors are cosine and proportion function. These two functions are symmetric similarity measures, which cannot find out subset copies. The asymmetric similarity model can detect subset copies. We develop two new asymmetric measures: HFM and HIPM. The HFM and HIPM are derived from cosine function and proportion function by combining asymmetric similarity concept with HFV. The HFV uses a short vector to represent a document in order to save resources and CPU time. We can adjust the dimensions of HFV to balance the model's performance. In brief, we can select appropriate similarity measure according to our environment and purpose.

Acknowledgements Our study is supported by national natural science foundation of china, ID: 60173058 (NSFC) and Xi'an Jiaotong University Science Research Fund (ID: 573031).

References

Brin S et al (1995) Copy detection mechanisms for digital documents. In: Proceedings of the ACM SIGMOD annual conference, San Francisco

Broder AZ et al (1997) Syntactic clustering of the Web. In: Proceedings of the 6th international Web Conference, Santa Clara

Denning PJ (1995) Editorial: plagiarism in the web. Commun ACM, 38(12)

Garcia-Molina H et al (1996) dSCAM: finding document copies across multiple databases. In: Proceedings of 4th International conference on parallel and distributed systems, Miami Beach

Heintze N (1996) Scalable document fingerprinting. In: Proceedings of the 2nd USENIX workshop on electronic commerce, Oakland

Monostori K et al (2000) MatchDetectReveal: finding overlapping and similar digital documents. In: Proceedings of information resources management association international conference, 21–24 May 2000 at Anchorage Hilton Hotel, Anchorage

Shivakumar N, Garcia-Molina H (1995) SCAM: a copy detection mechanism for digital documents. In: Proceedings of 2nd international conference in theory and practice of digital libraries, Austin

Shivakumar N, Garcia-Molina H (1998) Finding near-replicas of documents on the web. In: Proceedings of workshop on Web databases

Si A (1997) CHECK: a document plagiarism detection system. In: Proceedings of ACM symposium for applied computing, pp 70–77

Song QB, Shen JY (2001) On illegal coping and distributing detection mechanism for digital goods. J Comput Res Dev 38(1):121–125