Bulletin of Electrical Engineering and Informatics
Vol. 3, No. 4, December 2014, pp. 285~292
ISSN: 2089-3191

A Similarity Detection Method Based on Distance Matrix Model with Row-Column Order penalty Factor

Jun Li*, Yaqing Han, Yan Niu
Hubei University of Technology, Wuhan City, Hubei Province, China
*Corresponding author, email: 35113479@qq.com

Abstract

Paper detection involves multiple disciplines, and making a comprehensive and correct evaluation of academic misconduct is quite a complex and sensitive issue. There are some problems in the existing main detection models, such as incomplete segmentation preprocessing specification, impact of the semantic orders on detection, near-synonym evaluation, slow paper backtrack and so on. This paper presents a sentence-level paper similarity comparison model with segmentation preprocessing based on special identifier. This model integrates the characteristics of vector detection, hamming distance and the longest common substring and carries out detection specific to near-synonyms, word deletion and changes in word order by redefining distance matrix and adding ordinal measures, making sentence similarity detection in terms of semantics and backbone word segmentation more effective. Compared with the traditional paper similarity retrieval, the present method adopts modular-2 arithmetic with low computation. Paper detection method with reliability and high efficiency is of great academic significance in word segmentation, similarity detection and document summarization.

Keywords*: paper detection, segmentation, similarity comparison, distance matrix, model modular-2 arithmetic*

1. Introduction

With the development of network and information technology, information sharing and dissemination is becoming more and more convenient, providing a good platform for academic researchers, but at the same time, producing corresponding academic misconducts, which mainly displays in plagiarism of academic achievements of others. In addition to completely copy, one may plagiarize others' achievements by means of transposition, paraphrasing and synonym replacement. These behaviors cause damage to academic quality and in order to change this condition, paper similarity detection and other means are required to reduce occurrence of misconducts. Meanwhile, paper detection method with reliability and high efficiency is of great academic significance in word segmentation, similarity detection and document summarization.

Making a comprehensive and correct evaluation of academic misconduct and the cognizance of academic misconducts are quite complex and sensitive issues. Paper similarity detection is based on similarity calculation and uses computers to calculate the similarity between texts automatically. The calculation of text similarity has been widely applied in such fields as retrieval, machine translation, question answering systems and text mining, which is a basic and key problem and has long been a hot and difficult point for researchers. At present, many scholars both at home and abroad are studying text similarity calculation problem and put forward some solutions.

In 1993, Manber [1] from Arizona University proposed the concept of approximate exponential, which was the earliest detection technology and was used for measuring string similarity between documents. Later, Bao Junpeng [2] and others proposed document copy detection method based on semantic sequence, which emphasized position information of words and improved the detection accuracy, and afterwards they proposed corresponding detection model (SSK) [3], which was suitable for copy detection without replacement of words.

In 2005, Jin Bo [4] from Dalian University of Technology extended text similarity calculation based on semantic understanding into paragraph building on How Net semantic similarity, then extended paragraph similarity calculation into chapter and provided text

Received June 21, 2014; Revised October 4, 2014; Accepted October 28, 2014

(including words, sentences and paragraphs) similarity calculation formula and algorithms. In addition, he proposed long text similarity copy detection algorithms [5] in 2007, which calculated the coverage of similar semantic sequence set by semantic sequence similarity relation and chose each overlap values with the minimum entropy for plagiarism recognition and retrieval.

Hung Chenghui [6, 7] from Zhong Shan University and others proposed a text similarity calculation method by combining word semantic information with TF-IDF. In this method, sentence was seen as vector space composed of independent entry and the matching problem of document information was transformed into that of vector in the vector space and then the similarity was obtained by dot product method and cosine method. However, TF-IDF based on vector space model also has disadvantages. Firstly, it is a statistical-based method and only when the text includes enough words can some words occur repeatedly, which will reflect its statistical results. Secondly, this method only considers statistical attribute of words in context and ignores the semantic information, thus producing some limitations. Just as Salton form Cornell University says, there is no strict theoretical basis for calculating vector similarity by angle cosine.

To enhance the performance of result merging for distributed information retrieval, a novel merging method was put forward by Wang Xiuhong [8] from Jiang Su University and others, which was based on relevance between retrieved results and query.

In 2009, Nie Guihua [9] from Wuhan University of Technology proposed systematic frame model of ontology-based thesis copy detecting system, which described framework of paper copy detecting system from three layers : ontology access layer, ontology represent layer and ontology map layer. Semantic and ontology technology was utilized to discuss the build of paper ontology and calculation of paper similarity.

Zhang Huanjiong [10] from Beijing University of Posts and Telecommunications constructed a new formula to calculate text similarity from Hamming calculation formula based on Hamming distance theory. It had the advantage of simplicity and rapidity but ignored the effects of unequal-length text and word order.

Later in 2011, Chen Yao-Tsung [11] from Tai Wan proposed using chi-square statistics to measure similarities of term vector for texts, which reduced the miss rate of similarity detection.

This year, Sánchez-Vega [12] from Mexico proposes a rewriting exponential model based on finite state machine, which is able to detect common actions performed by plagiarists such as word deletion, insertion and transposition, but does not deal with near-synonyms.

Through above analysis, it can be seen that there exists problems in main detection methods at present as follows:

(1) The statistical property of words in the context is taken into consideration while semantic information is ignored.
(2) The effects of unequal-length text and word order are not well considered.
(3) There is limitation in looking up paper source by backtrack.

Chinese has the characteristics of big number in vocabulary and mutability in semantics, therefore, copy detection of Chinese paper is more complex and difficult than that of English paper. In addition, the basic resource (such as corpus) for Chinese language processing such as text detection is relatively lacking, making it impossible to apply some mature technology and research achievements abroad to Chinese paper detection directly. As a result, information processing of Chinese by computers is more difficult than that of western languages. Meanwhile, there are significant differences between Chinese and English segmentation because Chinese text is a continuous string of large character set, which means no specific separation mark exists between Chinese words, while English text is a fully separated string of small character set by space. Therefore, similarity comparison between Chinese texts must be based on segmentation systems, and the accuracy of word segmentation determines that of paper similarity calculation. Despite the achievements of word segmentation algorithm, there still exist problems as follows:

(1) Disunity in segmentation specification and inconsistency in dictionary design;
(2) Incompleteness of segmentation algorithm;
(3) Lack of reasonable ambiguity correction mechanism.

Word segmentation is the basis and premise of Chinese text similarity calculation, which means adoption of segmentation algorithm of high efficiency, can greatly improve the accuracy of text similarity calculation. Therefore, it is essential to keep on further exploration of new

segmentation algorithm on the basis of existing ones, and improve the integrity and accuracy of word segmentation to make the comparison of similarity between texts more accurate, thus providing decision support for related business.

On the other hand, similarity calculation, algorithm of which is often represented by formula or model of similarity calculation, has different requirements for different applications and has different levels such as paper-level, paragraph-level, sentence-level, word-level and morpheme-level. However, with the improvement of plagiarism means and variation of expressions, there exists such problems as large computation and difficulties in feature extraction and correct understanding of evaluation standards in paper-level and paragraph-level similarity detection, and as for word-level similarity, it may deviate greatly from actual value because of its small particle size. In addition, the space complexity and time complexity of detection algorithm are quite big, making it difficult to obtain ideal effect from its application into huge amounts of paper.

Starting with word segmentation and text similarity research, this paper takes advantages of several main detection models at present, builds the model specific to plagiarism, replacement of words and transposition based on the analysis of existing detection technologies and then presents a sentence-level paper similarity comparison model with segmentation preprocessing based on special identifier. This model integrates the characteristics of vector detection, hamming distance and the longest common substring and carries out detection specific to near-synonyms, word deletion and changes in word order by redefining distance matrix and adding ordinal measures, making sentence similarity detection in terms of semantics and backbone word segmentation more effective and the returned results of retrieval better meet the requirements of users. Compared with the traditional paper similarity retrieval, the present method adopts modular-2 arithmetic with low computation and its effectiveness and practicability are confirmed by corresponding experiment.

2. Distance Matrix Model with Row-Column Order Penalty Factor

2.1 Segmentation Based on Special Identifier and Indentification of New Words

In the process of text similarity calculation, accuracy of which determines that of paper similarity calculation, segmentation preprocessing must be carried out firstly. There are several common algorithms of Chinese segmentation at present, such as unsupervised segmentation, mechanical word segmentation based on dictionary, segmentation based on linguistic model and segmentation based on character tagging. The first algorithm judges the correlation between two characters by computing their mutual information in corpus, which has good effect with high frequency words but is affected by threshold coefficient; the second one does not have universality because it is affected by specialties of dictionary, and the third algorithm is the most commonly used method which is under research and development. There are two difficulties in Chinese Segmentation: ambiguity and identification of new words.

According to the recently published <<Chinese Grammar Questions>> (Another version of <<Chinese Grammar>>) of Professor Xing Fuyi, Chinese words are classified into three categories and eleven small classes, in which nouns, verbs, adjectives, numerals, quantifiers and pronouns are classified as notional words and adverbs, prepositions, conjunctions, auxiliary words, onomatopoeias and interjections are classified as functional words, which shows POS features of text can be used as basis of recognition of special identifiers.

From discussion above, it can be concluded that special identifiers are words or symbols of specific significance and function, which are composed of non-Chinese characters such as punctuations, unit symbols, mathematical symbols, numerical symbols and letters, and Chinese characters such as special symbols.

Firstly, <<Chinese Grammar>> is utilized to analyze the semantic and POS features of text, and then a special identifier dictionary is constructed to recognize special identifiers in the text and confirm them. After that, the number of characters in the word that will be segmented is determined. Match algorithm of two-character words is most widely used at present while that of multi-character words needs further improvement. To reduce number of queries is one of goals of this method, thus improving segmentation efficiency. The improved match algorithm of two-character words has such data structure for the segmented words as follows:

CiCi+1	N

In the table above, CiCi+1 is the segmented word and N is length of it. Based on this structure, each word CiCi+1… can be searched through its index Ci by building a hash table. According to the characteristics that two-word words are abundant in Chinese, match algorithm of two-character words is set as the main matching algorithm. Meanwhile, two concurrent processes are set in this system, one of which is used for searching the target two-character word in the corpus and if the search fails, add one character to the target word and continues until the length of the word is equal to that of the longest word in corpus, and in this case, the first character of the word will be segmented as a single word, while the other process is used for searching the target word in the dictionary composed of segmented words and if it succeeds, both processes are aborted, thus reducing the queries in the large corpus, otherwise the system keeps running until the first process succeeds, and then add the word into segmented dictionary. Through the combination of Baidu dictionary and the corpus, a following preliminary segmentation result can be obtained:

In Figure 1, '*'stands for special identifiers in Chinese.

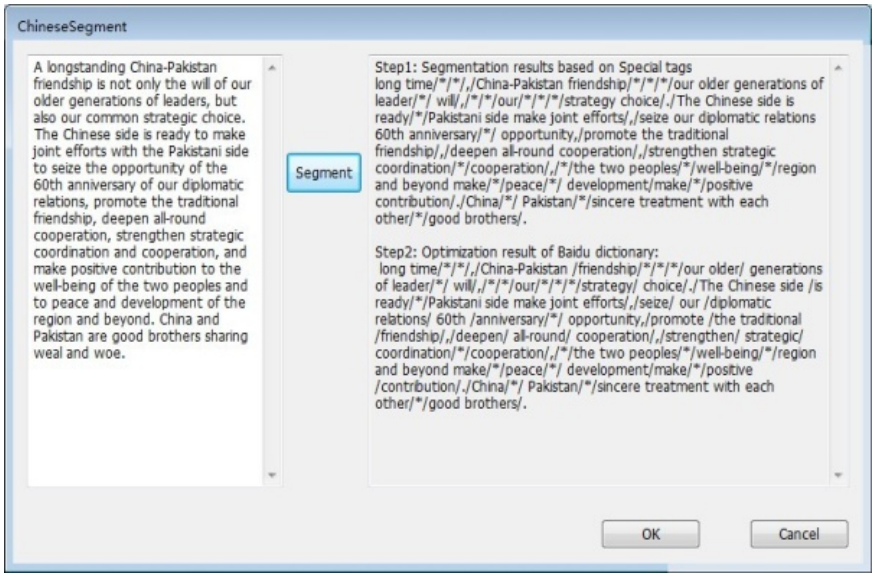

Figure 1. Segmentation result based on special identifier and Baidu Dictionary

2.2 Evaluation Score Model of New Words and Popular Words in Search Engine

The internet language develops at a high speed and new words appear continuously. Search engine has unique advantages in processing new words, popular words and common ambiguous segments because of its wide applications, and so assessment of new words and popular words can be conducted with the help of search engines' results. Based on special identifier, this paper searches the key words in search engine and regards the number of pages that are searched as their popularities. For instance, the search result of "Gangnam style" on Baidu is: "The number of related results is about 17,000,000", while that of "Gangnam styel" is: "The number of related results is about 1,790", which obviously shows that the former should be regarded as a whole word and added into the corpus. Through this method it can be judged whether a word is a new word or not. The algorithm of the popularity of a word is as follows:

$$P(a) = \frac{N}{100,000,000} \quad (1)$$

In (1), N is number of pages on Baidu and the denominator 100,000,000 is the maximum number of pages of searching the commonly used words on Baidu.

2.3 Research on Sentence-Level Similarity Algorithm

Automatic summarization is divided into key phrase extraction and text generation, which is used to explain and summarize the content of the original text with a brief description, accelerating people's understanding of information and solving the problem of information overloaded effectively. On the basis of statistics, paper-level and paragraph-level summary is composed of key sentences extracted from the text by Bayesian method and Hidden Markov Model according to word frequency and position information of the text, which is mainly suitable for technical documents with standardized format, and for other types of document, it needs further improvement in plagiarism-detection. Because of its unidirectional mapping mechanism, summary method is of good effect on complete paragraph plagiarism, which is rarely used in actual situation. In addition, it is of great complexity to extract key information with representativeness for paper-level and paragraph-level summary. However, the theory of summary can be used to create a digital signature which is composed of paper title and author information for each sentence and add it to sentence-level corpus for the subsequent backtrack.

A paper is composed of multiple sentences and after the segmentation is carried out, each sentence can be regarded as a set of words and then a sentence-level similarity comparison of the paper to be detected with the compared paper will be conducted. Firstly, add the segmentation results of compared paper into database, and secondly compare each sentence in the paper to be detected with that in the database. In order to look up the paper source by backtrack, a digital signature composed of paper title and author information is appended to each sentence, thus in the similarity comparison process, if two sentences are judged to have a high similarity, their corresponding papers can be retrieved quickly. This paper adopts MD5 algorithm, which can process messages of any length and produce a message summary with a fixed length, so the massage summary obtained through the input of paper title and author information is added into the sentence-level corpus as a digital signature. Sentence-level summary is shown as follows:

string	Objects Similarity Correlation Calculate Become Data Mining Information Extract Field Problem
title	Ontology Based Semantic Similarity and Relatedness Measures Review
author	Liu Hong-zhe Author
hash value	A8DDED5E696AE135F5C9D6EDFB016ADE

Figure 2. Result of sentence-level summary.

2.4 General Similarity Comparison Model with Sequence Factor

VSM (such as IF-IDF) transforms matching problems of document information to that of vector in the vector space with a comprehensive consideration of term frequency (TF) in all texts and the term' ability to distinguish catalogs (IDF). Hamming Distance adopts modular-2 arithmetic, avoids a mass of multiplication during similarity process in Euclid space, and obtains a high calculation speed. All these methods have their advantages, but at the same time, they have such problems as slow paper backtrack and no consideration for word order. For instance, in the following sentences: (a) "lexical collocation methods are various"; (b)" lexical types are various"; (c)" English lexical collocation methods are various", the similarities which are obtained by VSM, Hamming Distance and the longest common substring, respectively, are approximate if there no consideration of word order, but obviously, the similarity of sentence a and sentence c is higher than that of sentence a and b, so comparison of sentence-level similarities with consideration of word order will obtain better effects.

Distance matrix model with row-column order penalty factor in this paper integrates the characteristics of VSM, hamming distance and the longest common substring and realizes the transformation among these models through different configurations.

In this model, paper to be detected, called A, and compared paper, called B, are divided into two sets of sentences, and each sentence is a segmentation vector. So after the division, paper A is described as:
LA={A_1,A_2,A_3……A_n} ,paper B is described as: LB={B_1,B_2,B_3……B_n}, and if the two vectors have a different number of elements, add some empty string elements into the one with fewer elements until it has as many elements as the other, and then conduct the following calculations:

$$LA'\otimes LB = \begin{pmatrix} A_1 \\ A_2 \\ \cdots \\ A_n \end{pmatrix} \otimes (B_1 \quad B_2 \quad \cdots \quad B_n) \qquad Sim(LA', LB) = \frac{LA'\otimes LB}{k\sqrt{\sum_{i=1}^{n} A_i^2 * \sum_{j=1}^{n} B_j^2}} * f(TA, TB) \qquad (2)$$

$|LA'\otimes LB|$ is the maximum number of words that are the same in two sentences, which is similar to the longest common substring, and the difference is that the effects of multiple substrings that are the same is taken into consideration in this method. XOR is adopted in the calculation of $LA'\otimes LB$, result of which is an N-order sparse matrix. When the element from vector LA is multiplied by the element form vector LB, the result will be 1 if they are the same, otherwise it will be 0. The first non-zero elements in rows are counted to create a row-sequence table, called TA, saving the row numbers of those elements that have identical elements in LB, and a column-sequence table ,called TB, saving the column numbers of those elements that have identical elements in LA. TA and TB represent the word orders of the longest substrings in LA and LB, respectively. Then the similarity of TA and TB is compared by utilizing Euclidean distance and the following function $f(TA, TB)$ is called:

$$f(TA,TB) = sqrt(\sum_{i=1}^{n} |TA_i - TB_i|^2)$$

In which, i is the serial number of TA and TB, n is the elements number of TA and TB, and the result of the function will be regarded as the row-column order penalty factor. For instance:
Ps (compared sentence):

we guested Beijing visitors passionately
101 203 254 357 656

Pa (sentence to be detected):
Beijing visitors passionately guested we
Pb (sentence to be detected):

we passionately guested Beijing visitors

After segmentation, the sentence vectors are:

s=[101 * 203 * 254 * 357 * 656];
a=[254 * 357 * 656 * 203 * 101];
b=[101 * 656 * 203 * 254 * 357];

‘*’means there may be other words among the given ones. Through the vector multiplication and deletion of rows of all zero members, a matrix is obtained as follows:

$$a'\otimes s = \begin{bmatrix} 254 \\ 357 \\ 656 \\ 203 \\ 101 \end{bmatrix} \otimes \begin{bmatrix} 101 & 203 & 254 & 357 & 656 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

From the result, TA of sentence $a'\otimes s$ can be expressed as: (3, 4, 5, 2, 1) and similarly, TA of sentence $b'\otimes s$ can be expressed as: (1, 5, 2, 3, 4). With the assumption that word order of sentence Ps is: (1, 2, 3, 4, 5), the word order curve graph of three sentences can be obtained as Figure 3. And the Euclidean distances of Pa and Pb are as Figure 4 and Figure 5.

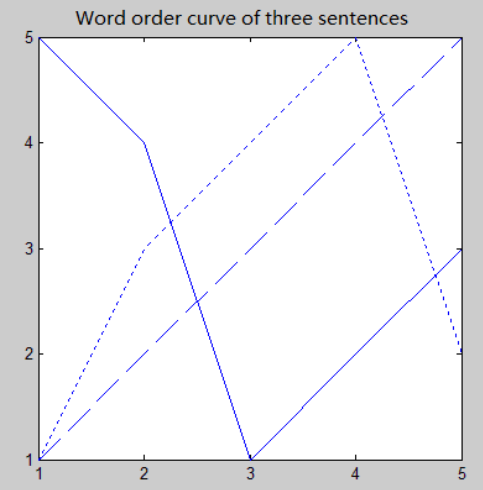

Figure 3. Word order of three sentences

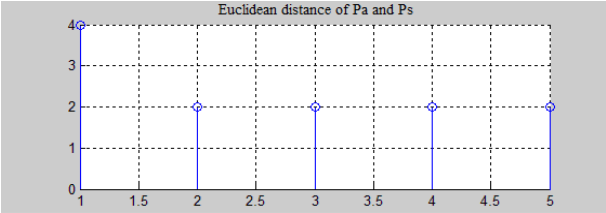

Figure 4. Euclidean distance of Pa and Ps

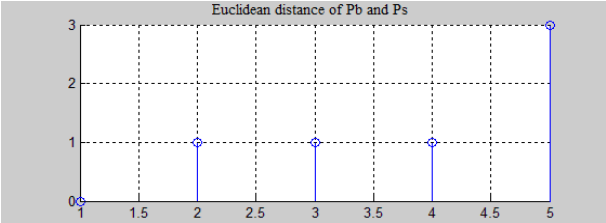

Figure 5. Euclidean distance of Pb and Ps

After calculation, the results of f(Pa, Ps) and f(Pb, Ps) are: 5.6569 and 3.4641, which shows penalty factor of Pb is smaller than that of Pa, meaning similarity between Pb and Ps is higher than that between Pa and Ps, which obviously agrees with manually assessing results, under the condition that the number of words that are the same between Pa and Ps equals to that between Pb and Ps. If $f(TA,TB)$ is set to 1, which means word order is ignored, then this model evolves into VSM model with Dice coefficient and if the weight of each word is ignored and multiplication is set as the algorithm, this model evolves into hamming distance model. In the model, k is a constant related to vector, if k is set to:

$$\frac{1}{\sqrt{\sum_{i=1}^{n} A_i^2 * \sum_{j=1}^{n} B_j^2}}$$

and the number of words that are exactly the same in the compared sentence and the detected sentence is defined as the rank, this model evolves into the longest common substring under the condition that word order is ignored, otherwise it is the longest common substring affected by order penalty factor.

2.5 Processing of Replacement of Words and Paraphrasing

For phenomenon of common replacement of words and paraphrasing, processing of near-synonym is adopted in this model, and the corresponding dictionary of near-synonyms, which are mainly from <<Chinese Synonyms Dictionary>> and those that are the same word in English, is build. In calculation of (2), the result of multiplying two near-synonyms is 1, the same as that of two identical words. For instance, “increase” and “Improve” are of the same meaning, if two sentences include “increase radiation” and “improve radiation”, respectively, the multiplication result of the two words will be 1 after they are judged to be near-synonyms through the dictionary of near-synonyms.

3. Conclusion

This paper proposes a segmentation algorithm based on special identifier and a new word recognition model based on search engine. Sentence-level similarity comparison model can evolve into three main similarity detection models: VSM, hamming distance and the longest common substring model, which obtains good effect on solving the paper plagiarism problem such as word deletion and addition, part copy and replacement of words by add order penalty factor and the process of handling near-synonyms. From the experiment result, it can be seen that the method proposed in this paper has a good effect on similarity comparison.

References

[1] Manber U and G Myers. Suffix arrays: A new method for on-line string searches. *SIAM Journal on Computing*. 1993; 22(5): 935-948.
[2] BAO Jun-Peng and SHEN Jun-Yi. A Survey on Natural Language Text Copy Detection. *Journal of Software*. 2003; (10).
[3] SHI Yi. Variance clustering based outline identification algorithm for time series date. *Journal of Computer Application*.
[4] JIN Bo, SHI Yan-Jun and TENG Hong-Fei, Similarity algorithm of text based on semantic understanding. *Journal of Da Lian University of Technology*. 2005; (02).
[5] FENG Zhong-Hui, BAO Jun-Peng and SHEN Jun-Yi. Incremental Algorithm of Text Soft Clustering. *Journal of Xi’an Jiaotong Universiry*. 2007; (04).
[6] HUANG Cheng-Hui, YIN Jian and HOU Fang, A Text Similarity Measurement Combining Word Semantic Information with IF-IDF Method. Chinese *Journal of Computers*. 2011; (05).
[7] HUANG Cheng-Hui, YIN Jian and LU Ji-Yuan, An improved Retrieve Algorithm Incorporated Semantic Similarity for Lucene. *Journal of Zhongshan University (natural science edition)*. 2011; (02).
[8] WANG Xiu-Hong and JU Shi-Guang, Result merging method based on combined kernels for distributed information retrieval. *Journal on Communications*. 2011(04).
[9] NIE Gui-Hua, Ontology-based Thesis Copy Detection System. *Computer Engineering*. 2009; (06).
[10] ZHANG Huan-Jiong, WANG Guo-Sheng and ZHONG Yi-Xin, Text Similarity Computing Based on Hamming Distance. *Computer Engineering and Application*. 2001; (19).
[11] Chen Y and MC Chen. Using chi-square statistics to measure similarities for text categorization. *Expert Systems with Applications*. 2011; 38(4): 3085-3090.
[12] Sánchez-Vega F, et al., Determining and characterizing the reused text for plagiarism detection. *Expert Systems with Applications*. 2013; 40(5): 1804-1813.