Current Trends in Technology and Science
ISSN : 2279-0535
8thSASTech 2014 Symposium on Advances in Science & Technology-Commission-IV Mashhad, Iran

Introducing an Automated Technique for Bilingual Plagiarism detection of English-Persian Documents

Soraya Enayati Shiraz
Master Computer Engineering, Islamic Azad University , Science and Research Branch of Tehran (Semnan),semnan, Iran, enayatishiraz@yahoo.com

Farzin Yaghmaee
Department of Electrical and Computer Engineering, University of Semnan, Semnan, Iran
f_yaghmaee@semnan.ac.ir

***Abstract* — Easy access that Internet has provided to vast quantities of electronic data, textual plagiarism has become a major concern especially in academic documents and research and scientific institutions. So with increasing rate of amount of information on the Internet, the problems and disorders resulting from bilingual plagiarism have resulted in solutions that they can be detected with auto-detection methods. The recommended detection methods are mostly used for bilingual plagiarism in English- Spanish, Bengali, German, French, and Vietnamese etc. In this paper, a method has been proposed that based on overall reliance of textual contents provides and using vector space model (VSM) can automatically detect bilingual plagiarism English-Persian. Method after implementation assessed using test texts through accuracy and recalling standards and F-criterion reliability and the results showed that the proposed method can detect English - Persian bilingual plagiarism with accuracy criteria of 0.88 and a reliability of 0.91.**

***Keyword* — Plagiarism, bilingual plagiarism detection, similarity analysis, morphological analysis, vector space model (VSM).**

1. INTRODUCTION

One of the fast and easy ways to access updated academic and research resources around the world is through Internet which along with its advantages, it has its own disadvantages as well. Including its disadvantages we can point out to easier stealing the scientific researchers' literatures by jobber people and this technique is also a growing challenge in the virtual world. Plagiarism means re-use of the ideas, results and/or the words of another person who has presented them for the first time without explicitly mentioning the references and authors [1,2]. "Textual" Plagiarism is one of the most common types of plagiarisms which mostly take place in universities and official organizations and today with the increasing amount of information they are detectable using automated and sophisticated methods. [2] In a general categorization, language text plagiarism detection is categorized into two categories: the first category is monolingual and/or homogenous English in comparison with English; the second category is cross-lingual or heterogeneous textual plagiarism detection such as English with Persian [3, 5].

Monolingual textual plagiarism detection techniques are mostly based on lexical, grammatical, semantic, structural characteristics of text and include methods based on textual comparison, discipline matching, writing style and fingerprint which are easily identifiable. [2,3,4]. But bilingual plagiarism detection methods have two levels, in first level the two languages have the same grammar and in second level the two languages do not have the same grammar. In case the grammars are not the same, the detection method becomes far more complex. Therefore texts classifications should be carried out in such a way that the words do not have dependency on the sentence level. Therefore bilingual text plagiarism detection is based on natural language processing techniques and machine learning techniques that in some of the methods text paragraphs are classified using some techniques and then with analysis on the paragraphs of the two texts and their similarities, the plagiarism can be recognized between them. In recent years, information retrieval techniques (based on word/Character n-gram (CNG), meaning-based, dictionary-based and bilingual website with ontology) [4,6,11,12], statistical methods [7], the method of support vector machine [8], statistical data analysis method and analysis of overall dependency of textual content [9,10] have been used. In the mentioned methods, the CNG-based method namely the method based on matching of the fields has a high accuracy but if the text volume is large or if the two languages do not have the the same grammar, it is not desirable.

The differences between English and Persian grammars, sentence structure and role of the words in a sentence, the problems in automatic translation from Persian to English and vice versa, Persian writing problems due to lack of standardized writing namely different uses for various forms of writing the words are the inherent problems of Persian language. Therefore, these problems have led to a condition in which texts translation is carried out based on text words and analysis of similarity based on overall dependence of the contents of the text.

In this paper a method has been proposed based on overall dependence of the text content [9, 10], using ineffective words removal, finding words roots with

morphology analysis by elimination of prefixes and suffixes of words and replacing the synonyms of the words, which converts the data text into a vector of words. Using the vector space model (VSM) [13] for classification and similarity analysis a system has been proposed which can detect bilingual (English - Persian) text plagiarism with a high precision.

2. THEGENERAL ARCHITECTURE OF THE PROPOSED METHOD

The proposed method is shown in Figure 1 with three main stages (database initialization and processing stage - storage stage - execution stage) and with sub-procedures and their relationships.

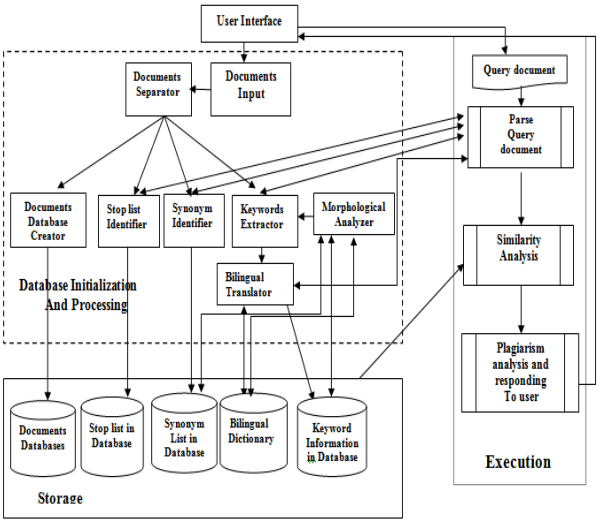

Fig.1. Overall architecture of bilingual plagiarism detection (English - Persian)

2.1. The database initialization and processing stage

1) *Input documents*: Input documents includes two categories: documents in English and documents in Persian, the list of ineffective words in both English and Persian, English and Persian suffix and prefix lists.
2) *Document Separator*: document separator separates input documents based on the contents of Persian documents collection and English documents collection and also performs the identification of stop word based on Persian and English text separation and identification of words synonyms.
3) *Documents database creator*: each collection of documents is saved in the documents database based on text information including ID, text title, and font size of each document.
4) *Stop list identifier*: At this stage ineffective words are detected and removed based on the list of ineffective words. Ineffective words are the frequent words which are repeated in most of the texts in order to convey the meaning of those texts. Due to the non-standard nature of Persian language in its written form, ineffective words have been considered in their different conditions of their writings. However, in recognition of combined words it can cause disruption in determining which form to use: considering the whole as a single word or as a few words but in English language, because of its standardized writing, it is much better.
5) *Synonyms identifier*: At this stage the synonyms of essential words are identified apart from ineffective words. Using Persian–Persian, English- English and Persian-English vocabularies and vice versa [14], the synonyms of required words are detected and stored in words database. But this step is one of the time consuming stages because some of the words have a different meaning in both Persian and English languages. Identification of words synonyms and accuracy in choosing alternative synonyms of words in both languages are effective in bilingual text plagiarism recognition accuracy.
6) *Morphological analyzer*: One of the text pre-processing stages is uniformization of text and the aim of uniformization is converting text to a uniform form. One of the uniformization methods is extraction of text characteristics which is possible through morphological analysis. In morphological analysis, each word is studied in terms of its presence in the vocabulary. If a word is not in the vocabulary, morphological rules are used to achieve the original word.Many of the words made up from one root word, have different meanings. Another problem is the words with different forms and different meanings derived from the same root. Namely they are different words in fact, however after the morphology they are converted to the equivalent words.
7) *Keywords extraction*: Keywords extraction is done using word stemming. In order to derive the root of words using morphological analysis, prefixes and suffixes removal procedure has been used. Identification of prefixes and suffixes of each language is possible through its grammar [15, 16] which is extracted as the output using the prefix and suffix lists and with searching and matching in the context of word roots.
8) *Bilingual translator*: with the help of bilingual dictionary, bilingual translator performs the necessary translation between the key words and checks if any of the keywords exists in the dictionary or not. If the keyword is not found in the dictionary of bilingual translator, it will insert it in the dictionary. And sometimes the meaning related to the language will also be inserted.

2.2. Storage stage

At this stage ,the information processed by database initialization and processing step is stored and also the mapping of key words are stored in bilingual dictionary in both Persian and English languages. The first phase is storage after collecting the initial data (input documents): Persian texts are as a text file of UTF-8 type and English

texts are as a text file of ANSI type and lists of ineffective words, synonyms, prefixes and suffixes are separated for each language.

In second phase of storage after processing, storing the key words of all English texts without duplicate words is carried out. Extracted key words from each text are stored as a vector of words based on the ID of each text (after morphological analysis). Storing is carried out after removing ineffective words and after applying synonym words according to the separation of each text in both languages.

2.3. Execution Stage

The process of execution stage has been carried out as shown in Figure 2:

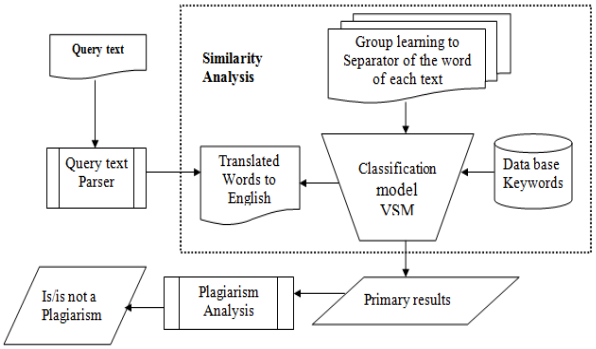

Fig. 2. The process of execution stage of the proposed method

1) *Receiving query text:* The query text in Persian language has been received from the user via system in a digital and automatically manner.
2) *Query text parser*: The query text parser first uses morphological analysis to find the root of some words and verbs by prefixes and suffixes removal method and then ineffective words are removed from the text and the synonyms of words essential words are detected and finally the extracted words of the text are translated by a bilingual dictionary. However, a preprocessing is required for standardizing the Persian text to convert non-standard forms to standard forms. Namely characters, punctuation marks and Persian words are written in the same form to be used in text analysis by computer. The result of this process is extraction of searched text features as a uniform form and in the form of features of educational texts which are the same as texts in English.
3) *Analysis of similarity*: To demonstrate the similarity of two texts and measuring the similarity criteria, vector space model (VSM) method has been used. This method is defined in a vector space. Using modeling text in vector space, vector space model considers a space vector for each text. The vector that each of its components is equivalent to a word from the entire text [8.13]. The general trend is as follows:

i. Classification training stage

i) Features Selection: This stage is selection of a subset of existing words in the English texts collection separated for each text. Text attributes have been created in initialization and processing stages based on overall dependency of the text content.

F_β	*Error Rate*	*Recall*	*Precision*	*Category*
0.91	0.08	0.96	0.88	Considering the morphology of synonyms and verbs.
0.77	0.21	0.90	0.68	Considering the Without morphology of synonyms and verbs.

ii) Weighting words: Assigning an appropriate numerical value to each of the selected features so differentiation of the text becomes more distinguishable than other texts.

Weighting words based on tf-idf[1] weighting has been performed based on equation (1) and equation (2) and equation (3).

$$d_j = (w_{1j}, w_{2j}, \ldots, w_{tj})$$
$$w_{ij} = tf_{ij}\, idf_i = tf_{ij} \log_2 (N/df_i) \quad (1)$$
$$tf_{ij} = f_{ij} / max_i\{f_{ij}\} \quad (2)$$
$$idf_i = \log_2 (N/df_i) \quad (3)$$

d_j: j-th text vector or j-th text properties vector, W_{ij}: the weight of i-th word in j-th text, f_{ij}: number of occurrences of i-th word in text j, df_i: number of documents in the training set in which the i-th word has occurred at least once, tf_i: represents the number of repetitions of the word i-th in the j-th text or the number of occurrences of each word, idf_i: Reverse of each text of the frequency of word i or inverse frequency of the text, N: the total number of documents

ii. Similarity measuring stage

Degree of similarity is calculated using cosine of similarity of each query text vector namely q = (w_{1q}, w_{2q}... w_{tq}) with resource vector set namely all the training texts using formula (4) [17]. And the value of similarity cosine will be reported in descending manner with mentioning the name of potential resource.

$$Cossim(d_j, q) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \cdot |\vec{q}|} = \frac{\sum_{i=1}^{t} (w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \cdot \sum_{i=1}^{t} w_{iq}^2} \quad (4)$$

iii. Plagiarism analysis and responding to user

By analyzing the test results obtained by testing many samples and calculating the amount of the cosine of similarity on plagiarized texts, a plagiarism level of 0.55 was considered. If the value of similarity cosine was greater than/or equal to 0.55 the text is plagiarized and if it becomes less than 0.55, the text is not a plagiarized one.

[1] Word Repetition in Reversed Text Repetition

3. Evaluating Results of the Proposed Method and Demonstrating Its Performance and Efficiency Styling

Evaluation data were collected from various Internet sources of bilingual texts and it was tried to choose them appropriately in terms of distribution. The evaluation data consisted of 100 training texts and 100 test text in two categories of texts belonging to training class and texts not belonging to training class and for each set, 50 test texts were considered. Experiments conducted on the selected texts were performed by applying synonym and morphological analysis and without morphological analysis and applying synonyms. System efficiency was evaluated according to precision, recall, error rates and F-measure using formula (5),(6),(7),(8) based on the following parameters. [18, 19]

$$\text{Precision} = \frac{TP}{\text{TP} + \text{FP}} \qquad (5)$$

$$\text{Recal} = \frac{TP}{\text{TP} + \text{FN}} \qquad (6)$$

$$F_{\beta} = \frac{(\beta^2 + 1) \times \text{Precision} \times \text{Recal}}{(\beta^2) \times \text{Precision} + \text{Recal}} \qquad (7)$$

$$\text{Error Rate} = \frac{FP + FN}{\text{N}} \qquad (8)$$

TP: the number samples which are member of class and properly detected, FP: the number samples which are member of class and not properly detected, FN: the number of class non-members and properly diagnosed, N: Number of texts tested, β: a parameter for controlling the balance between precision and recall which is assumed to be equal to 1. F_{β}: is the harmonic mean of precision and assessment and has been used to evaluate reliability.

According to the table, the results indicate that the system with the approach of applying morphology has a more precision than the approach without morphology and applying synonym in detection of English-Persian bilingual plagiarism. And also F_{β} criteria of 0.91 in fact show the high efficiency and reliability of the system.

Table (1) Assessment results obtained from testing 100 samples of Persian language text.

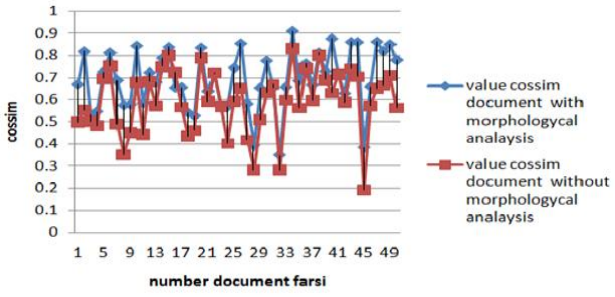

Fig. 3. Comparative value cossim test texts with morphological analysis and without morphological analysis

4. Conclusions

Plagiarism detection may play an important role in plagiarism of main ideas in papers, journals and internet websites. At present because of large volumes of digital information, it is impossible to evaluate them with manual methods. Thus, they must be recognized with automatic methods. Bilingual text plagiarism detection is based on natural language processing techniques and machine training methods. In this regard, information retrieval techniques (based on grammar- vector - meaning and a bilingual website with ontology) - statistical method- statistical data analysis method and general dependency analysis of textual content- support vector machine method have been presented for detection.

The proposed method is based on the overall dependence of the contents of text which can automatically recognize plagiarism in bilingual (English-Farsi) texts. According to the challenges in the field of translation from Persian to English and vice versa-lack of similar grammatical between English and Persian Languages and the non-standardized nature of Persian language writing method-text content analysis was carried out as words vectors at the overall text level and text features were selected based on overall dependency of text contents. In similarity analysis, creating correct correspondence between query text and the source texts according to information volume expansion is among the most important issues. In this method, VSM classifier has been used for text automated classification.

The proposed method is evaluated by testing Persian texts in two sets of plagiarized texts namely members of training class and non- plagiarized texts and non-member of training class with morphological analysis of verbs and words and without morphological analysis. Obtained results showed that the suggested method is very effective in detecting English-Persian bilingual text plagiarism. In cases where detection has not been correct (or were wrongfully detected), this can be attributed to text pre-processing, the method of writing, translation, and incorrect using the words synonyms and morphology of verbs.

5. Suggestions

Further research in order to automate the standardization of Persian texts writing, further research for better strengthening of Persian–English words synonyms automatic identification, and automatic texts translation through English - Persian bilingual translators as well as English - Persian bilingual text plagiarism detection with ontology are suggested as future studies.

Reference

[1] http://www.plagiarism.org
[2] Hermann Maurer, Frank Kappe, Bilal Zaka "Plagiarism - A Survey" Journal of Universal Computer Science, vol. 12, no. 8 ,2006.
[3] SalhaM.Alzahrani,NaomieSalim,andAjithAbraha, "Understanding Plagiarism Linguistic patterns

,Textual Features, and Detection Methods*"* IEEE Transaction Son System S, Man ,And cybernetics Partc : Applicationsandreviews,Vol.42,No.2,March 2012.

[4] Martin. Posthaste, Alberto Barron-Cedeno, Benno Stein, Paolo Rosso"cross-language plagiarism detection*"* lang resource& evaluation ,2010 .

[5] Asim M. El Tahir Ali,Hussam M. Dahwa Abdulla, and Vaclav Snasel " Overview And Comparison Of Plagiarism Detection Tools*"* ,2011.

[6] Chinh.Trong Nguyen, Dang Tuan Nguyen,"A New Model of English-Vietnamese Bilingual Information Retrieval System" World Academy of Science, Engineering and Technology 34,2009.

[7] A.Barron-Cedeno,P.Rosso,D.Pinto,andA.Juan,"On cross-lingual plagiarism analysis using a statistical model" in Proc. ECAI PAN Work-shop ,Patras , Greece, pp.9–13.

[8] Angel Anguita, Alejandra Beghelli And Werner Crelxell "automatic cross- language plagiarism Detection"detecting English-translated copy Spanish written documents 2011 IEEE.

[9] Mohammad shamsul arefine,yasuhliko morimoto, Mohammad amir sharif "bilingual plagiarism detector " international conference on computer and information technology(ICCIT2011)pp.22-24 December ,2011,Dhaka,Bangladesh.

[10] Mohammad shamsul arefine,et .al "BAENPD: A Bilingual Plagiarism Detector",Journal of computers, vol.8.NO.5,MAY ,2013.

[11] Rafal.CorezolaPereira,V.Moreira,andR.Galante,"*A new approach for cross-language plagiarism analysis*" in Multilingual and Multimodal Information Evaluation Access vol.6360,pp.15–26,2010.

[12] Rafal.Corezola pereira,"*cross- language tasks plagiarism detecation*," porto algre : program de pos-gradu ,computacao , 2010.

[13] V.Vapnik."The nature of statistical learning theory", 2nd edition, Springer, 2008.

[14] http://www.farsilookup.com

[15] http://www.fa.wiktionary.org

[16] Nagy, W.E., Berninger, V.W, & Abbott, R.D , "Contributions of morphology beyond phonology to literacy outcomes of upper elementary and middle-school students. *Journal of Educational Psychology*, *98*, 134–147, 2006.

[17] Sebastiani, F. "Machine learning automated text categorization", ACM Computing Surveys, 34(1), pp.1–47, 2002.

[18] S.Alvarez, "An exact analytical relation among recall,precision, and classification accuracy in information retrieval", Technical Report BCCS-02-01, Computer Science Department, Boston College, 2002.

[19] Y.Yang, " *An evaluation of Statistical Approaches to Text Categorization*", Kluwer Academic Publishers, Netherlands, 2000.