http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

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

Information Technology Journal 13 (14): 2349-2355, 2014
ISSN 1812-5638 / DOI: 10.3923/itj.2014.2349.2355

Arabic-English Cross-language Plagiarism Detection using Winnowing Algorithm

Adel Aljohani and Masnizah Mohd
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia,
Bangi, 43600, Selangor, Malaysia

Abstract: The availability of information in electronic forms and the availability of automatic translation machines has led to increased cross-language plagiarism. Manual detection of cross-language plagiarism is difficult, as such, developing an automatic system to detect such plagiarism is necessary. Although, there are a number of studies on detecting cross-language plagiarism in the form of Euro-English, Malay-English and Indonesian-English, there remains few studies concerned with the detection of Arabic-English cross-language plagiarism. This study proposes an Arabic-English cross-language plagiarism detection tool using the Winnowing algorithm. We evaluate its performance in terms of precision and recall on a data set consisting of Wikipedia articles. The performance of the proposed tool proved good with 97% precision, 81% recall and 89% F-measure evaluation metrics. The results show that the Winnowing algorithm can be used effectively to detect Arabic-English cross-language plagiarism.

Key words: Winnowing, plagiarism, Arabic, cross-language, Arabic-English plagiarism, cross-language plagiarism

INTRODUCTION

The term 'plagiarism' is of Latin origin and derives from the Latin word "plagiarius" which means kidnapper (Maurer *et al*., 2006). Plagiarism is an act of copying or attempting to copy or use the complete or partial content of another person's work's verbatim and failing to reference, cite or mention the original author of that content (Ali *et al*., 2011). There are numerous forms and approaches to plagiarism, such as:

- **Copy-paste plagiarism:** Direct verbatim copying, paraphrasing: Rephrasing certain content through different words
- **Translated plagiarism:** Content translation from one language to another and failing to recognize or reference the original study
- **Artistic plagiarism:** Presenting certain study through new mediums
- **Idea plagiarism:** The adoption of certain similar ideas which are not common knowledge and presenting them as one's own without reference to the original
- **Code plagiarism:** Which means the use of program codes without permission or reference from and to the original, absence of quotation marks which means the failure to recognize borrowed content replicated verbatim
- **Misinformation of references:** The inclusion of incorrect references (Maurer *et al*., 2006)

Plagiarism is categorized under electronic crimes like computer hacking, computer viruses, spamming, phishing and copyrights violation to name a few (Ali *et al*., 2011). There are three major techniques to automatically detect plagiarism, namely term occurrence, fingerprinting and style analysis (Anguita *et al*., 2011).

Detecting plagiarism is of particular importance in academia and the publishing industry as credibility in these institutions are largely based on originality (Khan *et al*., 2011). Plagiarism detection and prevention became one of the educational challenges, because most of the students or researchers are cheating when they do the assigned tasks and projects. This is due to the availability of the resources on the internet. It is easy to use one of the search engines to search for a specific topic and to cheat from it without citing the author of the document (Ali *et al*., 2011).

In this study, we present an Arabic-English cross-language plagiarism detection tool. We describe its main components including its pre-processing stage and the Winnowing algorithm. We evaluate it experimentally on a set of Arabic-English articles collected from Wikipedia.

Corresponding Author: Adel Aljohani, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, 43600, Selangor, Malaysia

BACKGROUND

Plagiarism: Plagiarism can be defined as submitting someone else's study as your own without reference to the original source (Lukashenko *et al*., 2007). Recent studies on plagiarism detection methods stated various practices of plagiarism including copying the whole or some parts of the document, rewording (paraphrasing and restating) the same content in different words, using others' ideas or referencing the study to incorrect or non-existing sources such as incorrect URLs or web pages that have been removed. Other practices of plagiarism include translated plagiarism (cross-language plagiarism) in which the content is translated and used without referencing the original study, artistic plagiarism in which different media such as images and videos are used to present other's study without proper citation and finally source code plagiarism (also called code clone) which can be defined as the reuse of the source code and similarly software designs and models without permission or citation (Maurer *et al*., 2006).

Plagiarism detection: Plagiarism can decrease through two methods, namely plagiarism prevention methods or plagiarism detection methods (Ali *et al*., 2011). Plagiarism prevention methods consist of two mechanisms, namely punishment routines and procedures pertaining to plagiarism drawback explanation. These methods are designed for a long-term positive effect, however, they require a considerable period in order to implement and harvest positive results as they depend on social cooperation between educational institutions. As for plagiarism detection methods, they are manual and electronic tools designed to identify plagiarism. However, they are largely seen as tools to address the problem and not the symptom and are not regarded as having a long-term effect. Despite this, when combined, both methods are an effective approach to reduce fraud and cheating.

There are three major techniques to automatically detect plagiarism, namely term occurrence, fingerprinting and style analysis (Anguita *et al*., 2011). The term occurrence technique rests on the assumption that similar documents consist of similar terms. As such it is possible to compare documents by evaluating the similarity in terms. The fingerprinting technique searches for a unique identifier (fingerprint) of a text. This unique identifier is then compared to other texts. When similar fingerprints are detected in other texts it is assumed to be similar and thus there is a good chance that the content was plagiarized. The style analysis technique examines the author, context and time of a text. From here, related texts the common features of other texts are codified under categories such as length of paragraphs and grammar. These common features are then used to compare with other texts to determine if there is content that matches the style of another text. This technique is popularly used to determine the authors of anonymous texts. Plagiarism detection is divided into two major classes, namely intrinsic plagiarism detection and external plagiarism detection (Potthast *et al*., 2009). Intrinsic plagiarism detection evaluates cases of plagiarism by searching into possible suspicious documents in isolation. This technique represents the ability of a person to detect plagiarism by examining differing writing styles. It seeks to identify potential plagiarism through the analysis of undeclared changes in writing style within a single document. External plagiarism detection assesses plagiarism in reference to one or more source documents in the data set. This process uses the ability of the computer to search for similar documents inside the corpus and retrieve possibly plagiarized documents. The typical process can be divided into the following three levels: Heuristic retrieval level, detailed analysis level and Knowledge-based post-processing level (Potthast *et al*., 2009). In heuristic retrieval level, a small set of documents is retrieved from the entire corpus. The retrieved set is likely to be the source of the query document. In the level of detailed analysis, the query document is compared (section-wise) to every retrieved document. Plagiarism suspicion parts and their potential sources are the identified. In the knowledge-based post-processing level, the short parts of text identified as plagiarized are discarded and identified neighbouring cases are merged to compose a single case.

Cross-language plagiarism detection: Based on language homogeneity or heterogeneity of the texts being compared, plagiarism detection can be classified into monolingual and cross-lingual (Alzahrani *et al*., 2011). The cross-language plagiarism detection process is similar to the external plagiarism detection task with some modifications in heuristic retrieval and detailed analysis stages (Barron-Cedeno, 2012). In cross-language heuristic retrieval, this stage aims to retrieve the collection of source candidate documents from the data set. Translating the input document from the query language to the source language may be required in this stage. The cross-language detailed analysis level measures the cross-language similarity between sections of the suspicious document and sections of the candidate documents which retrieved in the previous stage. There are five cross-language similarity analysis retrieval (Barron-Cedeno, 2012). Models based on syntax: This

model can be used to detect plagiarism between the languages with similar syntactic (English-French and Spanish-Catalan). Models based on thesauri: This model aims to bridge the language barrier by translating single words or concepts such as locations, dates and number expressions from L to L'. Models based on comparable corpora: In this case, the models for similarity assessment are trained over comparable corpora, i.e., a collection of documents C, C' where ci∈C covers the same topic than c'i∈C'. Models based on parallel corpora: In this case, the models for similarity assessment are trained over parallel corpora, i.e., a collection of documents C, C' where c∈C is a translation of c'∈C'. Models based on machine translation: This model is based on the principle of simplifying the problem by making it monolingual (Barron-Cedeno, 2012).

Fingerprinting technique: K-grams are central to fingerprinting techniques because fingerprinting divides the document into grams of certain length k (Zini *et al*., 2006). This allows the fingerprints of two documents to be compared in order to detect plagiarism. The fingerprint matching approach differs based on the comparison unit (i.e., grams). This technique can be classified into two categories, namely full and selective fingerprinting (Heintze, 1996).

Full fingerprinting: A full fingerprint of the document consists of the set of all the possible sequential substrings of length n in words or characters. There are |D|- n+1 substrings, where |D| is the length of the document in words or characters. This fingerprinting technique selects overlapping sub-strings.

Selective fingerprinting: There are various versions of selective fingerprints including "i^{th} hash", "0 mod P" and "Winnowing Algorithm" to decrease the size of a full fingerprint (Schleimer *et al*., 2003). In "i^{th} hash", every ith hash of a document will selected. This method is very easy to implement but has a poor result in case of insertion, deletion or reordering. For example, if the user adds only one letter into the text then the text fingerprints will shift by one which makes changes between the original and suspicious document's fingerprints, resulting in failure to detect plagiarism (Schleimer *et al*., 2003). In the "0 mod P" scheme, p is an integer and the hashes located at every "0 mod P" are selected. This method is very easy to implement but it is weak in terms of plagiarism detection cases (Schleimer *et al*., 2003). Three least frequent 4-grams is a selective fingerprinting technique that depends on calculating the weights of all 4-grams in the sentence then choosing the three least frequent 4-grams as sentence fingerprint (Yerra and Ng, 2005).

Winnowing algorithm: The winnowing algorithm is an algorithm to select document fingerprints from hashes of k-grams (Schleimer *et al*., 2003). To obtain the fingerprint of a document, the text is divided into k-grams, the hash value of each k-gram is calculated and a subset of these values is selected to be the fingerprint of the document. The example below shows the steps to get the fingerprint for the text "Kuala Lumpur".

The first step in the Winnowing algorithm is to remove irrelevant information from the text (whitespaces, punctuation, symbols).

- Step 1: Remove irrelevant features

Kualalumpur

In the second step, we create the k-gram with k length. The k length impacts on the efficiency of the algorithm where the big size of k can avoid the false positive cases but the algorithm will be insensitive for some plagiarism cases such as in words reordering and sentence restructure. In our example, we use the length 5.

- Step 2: Create 5-grams sequence

kuala ualal alalu lalum alump lumpu umpur

In the third step, the hash value for every 5-grams is calculated. The hash value will change the alphabet to integers.

- Step 3: Calculating the hash for every k-grams

18 22 20 23 50 44 24

In the fourth step, the window with size (w) is created from the hash values obtained in step 3.

- Step 4: Creating overlapping windows of length w (here we use w = 4)

(18 , 22 , 20 , 23)
(22 , 20 , 23 , 50)
(20 , 23 , 50 , 44)
(23 , 50 , 44 , 24)

In the next step, the smallest hash value in every window will be chosen and if there are two hashes with the smallest value, the rightmost will be chosen.

- Step 5: Choosing the smallest hash value from every window

(18 20 23)

It is useful to record in addition to the fingerprints of a document, where they are located within the document. For example, positional information is required to demonstrate the matching substrings in a user interface. An efficient implementation of the Winnowing algorithm must retain the position of the most recently selected fingerprint. The last step registers the fingerprint with its position (the first position is numbered 0).

- Step 6: The fingerprint selected by Winnowing with the 0-base positional information.

[18, 0] [20, 2] [23, 3]

To compare two fingerprints, the resemblance is calculated according to the following equation:

$$R=\frac{F(A)\cap F(B)}{F(A)\cup F(B)}$$

where, $F(A) \cap F(B)$ is the common hashes in sentence A and B, $F(A) \cup F(B)$ is the total number of all hashes in sentences A and B. If the value of the resemblance (R) between two fingerprints more than a predefined threshold, two fingerprints are similar. The threshold can be set depending on the desired check; threshold value should be small if we are looking for plagiarized parts like paragraphs or sentences while threshold value should be large if we want to test if documents share large content (Brin *et al*., 1995).

Hashing function: Comparison between strings is computationally and spatially expensive (Barron-Cedeno, 2012). As a result, models have been designed to represent text contents that require low amounts of space and make for an efficient comparison. This is precisely the case of the family of hash models. The purpose of a hash function is mapping a string (for instance, a word, a sentence, or an entire document), into a numerical value. In order to speed up the comparison process among documents, n-grams, sentences or fixed length text fragments from a document collection can be hashed. When analysing a suspicious document it can be hashed by means of the same function and queried against other hashes. If a match occurs, two exact text fragments have been found. In plagiarism detection cases, the advantages of the hash functions are the following: The resulting hash value is a compact representation of the input string, saving space and collisions are extremely unlikely (a collision implies obtaining the same hash value from two different text strings) (Barron-Cedeno, 2012).

There are many hash functions used by researchers in plagiarism detection studies, such as Md5, karp-rabin and BKDR. In our tool, the BKDR hash function was used to represent the text content in small numerical representation and reduce the memory space in comparison levels.

Measure of performance: There are two objectives in retrieval task. The first one is to retrieve the most of relevant documents and the second objective is to retrieve the least of irrelevant documents (Barron-Cedeno, 2012). Two classical measures in IR which aim at estimating how well these objectives are achieved are the well-known recall and precision metrics. A plagiarism detection system can be evaluated as a classification system where each sentence belongs to one of the two classes: Plagiarized or original. The result of detecting in plagiarism detection system can be divided into four types: True positive, true negative, false positive and false negative (Jadalla and Elnagar, 2012). True Positive (TP) is the set of plagiarized parts already detected by the system. True Negative (TN) is the set of non-plagiarized parts and the system selects them as such. False Positive (FP) is the set of non-plagiarized parts butthe system detected it as plagiarized. False Negative (FN) is the set of plagiarized parts butthe system did not detect it. In terms of these four sets, recall can be defined as follows:

Recall measure is defined as the percentage of relevant plagiarized parts detected by the system. Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

The second performance metric is the precision. Precision metric is used to measure the accuracy of the plagiarism detection system. Results indicate the percentage of plagiarism correctly detected by the system. The precision is defined as:

$$\text{Precision}=\frac{TP}{TP+FP}$$

Precision (positive predictive) is the part of retrieved documents that are relevant whereasrecall (sensitivity) is the part of relevant documents that are retrieved from the corpus. The high precision value refers to the effectiveness and efficiency, while the high recall value refers to the durability (Jadalla and Elnagar, 2012).

Both precision and recall are therefore based on an understanding and measure ofrelevance. As example,

search engine returns 100 pages, 50 of the retrieved pages were relevant while the search engine failing to return 100 additional relevant pages. The precision of the search engine is 50/100 = 0.5. The value of recall is 50/150 = 0.33. High recall means that the algorithm returned most of the relevant results, while high precision means that the algorithm already returned more relevant results than irrelevant.

The third metric is F-measure. F-measure combines precision and recall into a single measurement to balance them.

The range of F -measure is between 0 and 1. A combination of both measures (recall and precision) offers a better picture of an obtained result (Barron-Cedeno, 2012).

METHODOLOGY

Figure 1 depicts the overall processes and components of the proposed tool. The proposed tool consists of five stages. The first stage is the translating of input text from Arabic to English then the text pre-processing which consists of sentences identification, tokenization, stop-words removing and word stemming. The goal of the second stage is to convert the input text to fingerprint using the Winnowing algorithm. Winnowing algorithm used to reduce the index size and speeding processing time. The third stage aims at retrieving the most similar fingerprint to the input fingerprint from the fingerprint index. The goal of the fourth stage is to compare the input fingerprint with the retrieved fingerprint to detect the similarity. The final stage aims to show the detection result.

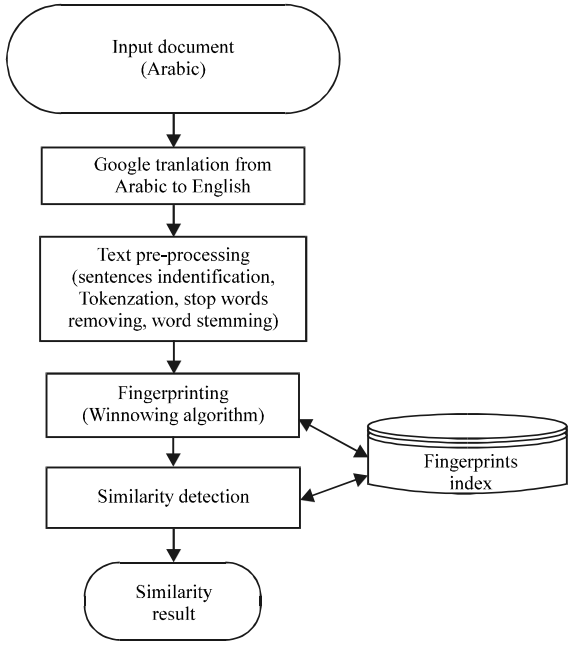

Fig. 1: Cross-language plagiarism detection tool processes

Google translation from Arabic to English: In order to detect cross-language plagiarisms it is essential to translate the input document from Arabic into English before used as the query document for further detection process. After the plagiarized documents have been translated into English it will improve the effectiveness of the detection process as the source documents are also in English. We use Google translate API which is a well-known translation tool developed by Google. With this API, the language blocks of text that can be easily translated to other preferred languages.

Text pre-processing and filtering: To improve the process of plagiarism detection results and to reduce the processing time, our corpus data needs to follow:

Sentences identification: Text segmentation in general is the process of dividing a text into local coherent clauses or sentences (text segments). In the plagiarism detection problem that we tackled in this study, sentence segmentation is not a major concern. The chosen delimiting punctuation marks were (.) , (!) and (?) where a sentence length should be no less than 8 words.

Tokenization: Tokenization is the process that splits the text into tokens. This process helps system to process each token separately and use them for other pre-processing steps, namely stop words removing and word stemming. Tokenization is easy to process in English as it splits tokens by white-spaces.

Removing stop words: Before passing the translated documents for comparison against the retrieved documents it is essential for us to remove the stop words in the translated text. English stop words will be removed in the translated texts. Currently, there are several English stop words commonly used in the information retrieval process. Some of the general English stop words include a, an, the, ourselves, been, anywhere, any, by, did, each, ever, even, would, could, few, than and all.

Word stemming: We use Porter stemmer in word stemming process. The Porter stemming algorithm is a process for removing the commoner morphological and inflectional endings from words in English. Its main use is as part of a term normalization process that is usually

done when setting up information retrieval systems. Porter stemmer is widely used as a stemming algorithm that is fully tested for its accuracy and effectiveness (Kent and Salim, 2009).

Implementing winnowing algorithm: To select the fingerprint for every sentence we implemented the Winnowing algorithm. To use the Winnowing algorithm, there are two parameters that must be chosen carefully, namely n-gram length (N) and window size (W). Both N and W can be empirically defined (Barron-Cedeno, 2012). Using very big or very small numbers for (N) and (W) will decrease the accuracy of the detecting process. The big number of N and W will increase the false negative while the small number of N and W will increase the false positive. We decided to use N = 5 and W = 4.

Building fingerprints index: To retrieve the most similar sentences from the corpus we implement an inverted index such as that used in search engines. Inverted index is a data structure that maps words to their locations in a collection of documents. To use the same structure in our study we have to implement two modifications in an inverted index design as used by Jadalla and Elnagar (2012). The first modification in our tool is a single fingerprint not a single word or term. The second modification, the basic form of an inverted list in the search engine is the set of documents that contain the index term, whereas in our tool, the inverted list consist of the set of sentences that contain that fingerprint. After that, every sentence in the input document is tested against possible plagiarism by one query. This query is made up of the whole set of fingerprints of the input sentence separated by the Boolean operator "OR". The search engine will then return the similar sentences to our query sentence.

Similarity detection: To compare the similarity between the fingerprint of the suspected sentence and the retrieved sentences from the fingerprint index, we use the resemblance measure according to the following equation:

$$R = \frac{F(A) \cap F(B)}{F(A) \cup F(B)}$$

where, $F(A) \cap F(B)$ is the common hashes in sentences A and B, $F(A) \cup F(B)$ is the total number of all hashes in sentences A and B. The result will be between 0 and 1, 0 means the two sentences are completely different whereas 1 means the two sentences are completely the same. In this study, if the resemblance equal or more than 0.5, the sentence is consider plagiarized.

RESULTS

Three main experiments were conducted to evaluate the performance of the proposed tool, statement-to-statement experiment, one-to-one experiment and one-to-all experiment. All plagiarized sentences were detected manually and by means of the detection tool. In statement-to-statement experiment, one Arabic sentence compared against one English sentence. In one-to-one experiment, one Arabic document was compared against one English document. In the one-to-all experiment, one Arabic document was compared against the whole data set. The performance results were measured using recall, precision and F-measure metrics.

Table 1 present the overall performance of the proposed tool through the conducted experiments whereas Table 2 shows the evaluation of results of the proposed tool. TP indicates true positive cases, while FP indicates false positive cases and FN indicates false negative cases. The proposed tool detected most of the plagiarised cases with little false positive cases.

The experiments show that the proposed tool using the Winnowing algorithm has a capability to detect Arabic-English cross-language plagiarism. The errors in automatic translation of the names (persons, places, companies and cities) from Arabic to English influenced the accuracy of the detection which decreases the accuracy of the tool. Using acronyms such as NLP, OOP and AI in the source documents has little influence on the decreased ability of detecting plagiarism where these abbreviations have no Arabic translation. Extracting these acronyms may decrease the false negative cases.

To compare the accuracy of the Winnowing algorithm in Arabic-English cross-language plagiarism detection against another plagiarism detection techniques we implement another plagiarism detection tool using the three least frequent 4-grams algorithm. The same experiments were conducted using this algorithm. Table 3 shows the overall performance of the tool using

Table 1: Overall performance of the proposed tool

No. of plagiarised sentences	Detected	TP	FP	FN
1403	1168	1142	26	261

Table 2: Evaluation of the results of the proposed tool

Evaluation	Result
Precision	0.97
Recall	0.81
F-measure	0.89

Table 3: Overall performance of the tool using the 3 least frequent 4-grams

No. of plagiarised sentences	Detected	TP	FP	FN
1403	981	942	39	448

Table 4: Evaluation of the results of the tool using 3 least frequent 4-grams

Evaluation	Result
Precision	0.96
Recall	0.67

three least frequent 4-grams algorithm whereas Table 4 shows the evaluation of the tool that used the three least frequent 4-grams.

The results shows that the Winnowing algorithm outperforms three least frequent 4-grams algorithm in detecting most Arabic-English cross-language plagiarism cases.

CONCLUSION

This study reviewed the problem of cross-language plagiarism and explored the existing plagiarism detection techniques. We have presented the first Arabic-English cross-language plagiarism detection to detect the Arabic sentences translated from English sources without mention of the original sources, in addition to describing its main components and processes.

Finally, we have presented and discussed the experiments conducted to demonstrate its effectiveness on a large set of Arabic-English articles. The result shows that the Winnowing algorithm can be used effectively to detect the Arabic-English cross-language plagiarism with 81% recall, 97% precision and 89% F-measure.

Future studies include testing the proposed tool on both intra-corpus (data set) and inter-corpus (world wide web) to compare the accuracy of detecting plagiarism in two cases. Future studies also include using different automatic translations such as Bing translation to compare with the results from Google translation.

REFERENCES

Ali, A.E.T., H.D. Abdulla and V. Snasel, 2011. Survey of plagiarism detection methods. Proceedings of the 5th Asia Modelling Symposium, May 24-26, 2011, Manila, Philippines, pp: 39-42.

Alzahrani, S.M., N. Salim and A. Abraham, 2011. Understanding plagiarism linguistic patterns, textual features and detection methods. IEEE Trans. Syst. Man Cybernetics C: Appl. Rev., 42: 133-149.

Anguita, A., A. Beghelli and W. Creixell, 2011. Automatic cross-language plagiarism detection. Proceedings of the 7th International Conference on Natural Language Processing and Knowledge Engineering, November 27-29, 2011, Tokushima, Japan, pp: 173-176.

Barron-Cedeno, A., 2012. On the mono-and cross-language detection of text reuse and plagiarism. Ph.D. Thesis, Universitat Politecnica de Valencia, Valencia, Spain.

Brin, S., J. Davis and H. Garcia-Molina, 1995. Copy detection mechanisms for digital documents. ACM SIGMOD Rec., 24: 398-409.

Heintze, N., 1996. Scalable document fingerprinting. Proceedings of the USENIX Workshop on Electronic Commerce, Volume 3, November 18-21, 1996, Oakland, CA., USA.

Jadalla, A. and A. Elnagar, 2012. A fingerprinting-based plagiarism detection system for Arabic text-based documents. Proceedings of the 8th International Conference on Computing Technology and Information Management, April 24-26, 2012, Seoul, Korea, pp: 477-482.

Kent, C.K. and N. Salim, 2009. Web based cross language semantic plagiarism detection. J. Comput., 1: 39-43.

Khan, M.A., A. Aleem, A. Wahab and M.N. Khan, 2011. Copy detection in Urdu language documents using n-grams model. Proceedings of the International Conference on Computer Networks and Information Technology, July 11-13, 2011, Abbottabad, Pakistan, pp: 263-266.

Lukashenko, R., V. Graudina and J. Grundspenkis, 2007. Computer-based plagiarism detection methods and tools: An overview. Proceedings of the International Conference on Computer Systems and Technologies, June 14-15, 2007, Rousse, Bulgaria.

Maurer, H., F. Kappe and B. Zaka, 2006. Plagiarism-A survey. J. Univ. Comput. Sci., 12: 1050-1084.

Potthast, M., B. Stein, A. Eiselt, A. Barron-Cedeno and P. Rosso, 2009. Overview of the 1st international competition on plagiarism detection. Proceedings of the SEPLN Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse, September 2009, Donostia-San Sebastian, Spain, pp: 1-9.

Schleimer, S., D.S. Wilkerson and A. Aiken, 2003. Winnowing: Local algorithms for document fingerprinting. Proceedings of the ACM SIGMOD International Conference on Management of Data, June 9-12, 2003, San Diego, California, USA., pp: 76-85..

Yerra, R. and Y.K. Ng, 2005. A Sentence-Based Copy Detection Approach for Web Documents. In: Fuzzy Systems and Knowledge Discovery, Wang, L. and Y. Jin (Eds.). Springer-Verlag, Berlin, Heidelberg, pp: 557-570.

Zini, M., M. Fabbri, M. Moneglia and A. Panunzi, 2006. Plagiarism detection through multilevel text comparison. Proceedings of the 2nd International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution, December 13-15, 2006, Leeds, UK., pp: 181-185.