Journal of Theoretical and Applied Information Technology
31st October 2011. Vol. 32 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CONCEPTUAL SIMILARITY AND GRAPH-BASED METHOD FOR PLAGIARISM DETECTION

[1]AHMED HAMZA OSMAN, [2]NAOMIE SALIM, [3]MOHAMMED SALEM BINWAHLAN, [2]HAMZA HENTABLY, [1,2]ALBARAA M. ALI

[1,]Faculty of Computer Science, International University of Africa, Sudan-81310

[1,2] Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Malaysia-81310

[3]Faculty of Applied Sciences, Hadhramout University of Science & Technology, Yemen-81310

E-mail: ahmedagraa@hotmail.com, naome@utm.my, moham2007med@yahoo.com, hentably@yahoo.com

ABSTRACT

Plagiarism is a form of academic misconduct. It has increased rapidly because it is now quick and easy to reach data and information through electronic documents and the Internet. The problem occurs when found documents content is illegal and without permission or citation, this problem is known as plagiarism. One of the major challenges is to detect the plagiarism and illegal copy. This paper discusses a new representation method for text documents called text graph-based representation. The proposed method does not represent the content of a text document as a graph only, but also captures the underlying semantic meaning in terms of the relationships among its concepts in order to defeat the difficulty which the traditional plagiarism detection systems face with some kinds of plagiarism such as complicated plagiarism in which users can reword the plagiarized part or replace some words by their synonyms. The experiments have been carried out using PAN-PC-09 standardization of plagiarism detection corpus. The results showed that our method remarkably outperforms the modern methods for plagiarism detection.

Keywords: *Concept Extraction, Graph Representation, Plagiarism Detection, Topic Signature*

1. INTRODUCTION

Nowadays, many resource documents are available in the internet and easy to access. Due to this availability, users can easily create a new document through copying and pasting from this resource. Sometimes users can reword the plagiarized part or : replace some words by their synonyms, where this kind of plagiarism is difficult to detect by the traditional plagiarism detection systems such as COPS, SCAM, MDR, etc. In many situations, plagiarists may not copy or change all the text, but they take relevant parts thinking it undiscovered plagiarism that can help their work. In this paper, conceptual meaning of the sentences has been semantically focussed. Most of tools used string matching algorithms to detect the plagiarism and ignored semantic matching between the similar documents. If the semantic change or paraphrasing of the text occurs, the detecting process will become difficult. Owing to the plagiarism matter, organizations have been built tools to avoid dishonest works. The challenge is to provide plagiarism checking technique with an appropriate algorithm in order to detect lexical and linguistic matching. In addition, to improve the percentage of finding result and time checking. But, the bigger challenge is how to prevent or minimize the technical issues such as unnecessary repetition. All of these reasons led us to rethink about a suitable method to handle plagiarism problems, and hence our proposed method comes as a solution for these problems.

In this paper, we propose a plagiarism detection system based on graph-based representation. For this purpose we used both content (words) and semantics to map textual sentences into semantic role structure. We think this is one of an interesting research problem, since automatic plagiarism detection would be very helpful for mentors, publishers, etc. to minimize the copy-paste problem which results in copyright violation and other

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ethical issues. We adopt the graph structure to represent the document, and then we use this representation in plagiarism detection by using semantic graph matching. The proposed method focus on solving the copy paste plagiarism detection and rewording using synonymy replacement. Many techniques such as [7] [34] [35] focused on the lexical plagiarism detection and ignored the semantic plagiarism. One of the main objectives of this paper is to capture the plagiarism semantically. for graph building, each sentence is represented as vertices, and relations between them are represented as arcs. All nodes have a direct edge with a unique node called Topic Signature Node. The similarity between the original and suspected documents is calculated through this node. The aspects of this paper seem as very interesting issues. We will see in the coming sections how these aspects lead us to avoid the plagiarism in general and reduce the violation chances of the authors copyright in particular.

The rest of this paper organised as follow: Section 2 provides a description of the related work in plagiarism detection and graph based representation. In Section 3, a description of the underlying idea of graph representation for plagiarism detection that involved in our method was covered. Section 4 discusses a methodology design and plagiarism detection based on our graph representation method. The experimental design and results evaluation of the proposed approach are introduced in section 5 whereas section 6 concludes the paper.

2. RELATED WORK

In this section, we review the fundamental concept and methods used in this paper. This covers two aspects, plagiarism detection and graph based representation.

2.1. PLAGIARISM DETECTION

In plagiarism detection, practical steps were proposed by Mallon[35], Martin [37], LaFollette [20], Hannabuss [20] and Angélil-Carter [5]. Joy [24] defined plagiarism as "unacknowledged copying of documents or programs". It can occur in many sectors, for example, companies may look for competitive advantage, and academicians need to advance their institutions by searching for quick ways for publishing. Most empirical studies and analysis were undertaken by the academic community to deal with student plagiarism. In order to discriminate plagiarized documents from non-plagiarized documents, a correct selection of text features is a key aspect. Clough [12] demarcates a set of features which can be used to find plagiarism aspects such as changes in the vocabulary, amount of similarity among texts or frequency of words. These features have produced different approaches to these aspects. Substantive plagiarism analysis is a different task from plagiarism detection with reference Meyer [28] . It captures the style across a suspected document in order to find fragments that are plagiarism candidates. This approach saves the cost of the comparison process, but it does not give any hint about the possible source of the potential plagiarized text fragments, which the search process has been based on. Lyon [23] considers text comparison based on word n-grams. With reference to this, the suspected text is split into tri-grams composing of two sets to be compared. The amount of common tri-grams is considered in order to detect potential plagiarism cases. Kang [16] considers the sentence as the comparison unit in order to compare local similarities. It differentiates among exact copy of sentences, word insertion, word removal and rewording. Several techniques have been developed or adapted for plagiarism detection in natural language documents. They can be classified into some number of main approaches. One technique is a Fingerprint Matching [15], [24], and [39]. It involves the process of scanning and examining the fingerprints of two documents in order to detect plagiarism. Clustering is another approach Antonio [6], Manuel[25] that uses specific words (or keywords) to find similar clusters between documents. Fingerprinting techniques mostly rely on the use of K-grams Manuel [25] because the process of fingerprinting divides the document into grams of k-lengths. Then, the fingerprints of the two documents can be compared in order to detect plagiarism. Fingerprints can be classified into three categories: character-based fingerprints, phrase-based fingerprints and statement-based fingerprints. The early fingerprinting technique uses sequence of characters to form the fingerprint for the whole document. Therefore, the proposed model significantly improves short coming of the existing plagiarism detection techniques.

Brin and Garcia-Molina [7] introduced plagiarism detection system from Stanford Digital Library Project named COPS (copy protection system), which detects document overlap relying on string matching and sentences. But, its main drawback is that it fails to consider individual words and takes the whole sentence as one part. The shortness of COPS was solved by Shivakumar and Garcia-Molian [34], who developed a new method

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

called Stanford Copy Analysis Method (SCAM) to improve the COPS. The SCAM used Relative Frequency Model (RFM) to stand out subset copies. RFM is an essential asymmetric similarity measure for plagiarism detection. The main advantage of SCAM is that, it can find the overlapping similarity between the part of sentences, but many terms misleads in documents sharing comparison. [35] Proposed a new mechanism for plagiarism detection called (CHECK), which is similar to SCAM. Both of them adopted information retrieval techniques and work for overlapping detection based on frequency of word. The CHECK technique, built on indexed structure known as structural characteristic (SC), is used to parse documents for building the SC. It captures the plagiarism, depending on the key words proportion of structural characteristic for the nodes. The limitation of the CHECK covered the structured documents only, where unstructured documents were ignored.

Match Detect Retrieval (MDR) was proposed by Krisztian [18]. In this system, plagiarism can be detected using string matching similarity algorithms based on suffix trees. The advantages of MDR concentrated in the copy paste plagiarism, but the limitation appearing when the plagiarized parts modified by rewording or synonyms replacement. Another limitation when we want to build the suffix tree for the suspected documents. Were the constructing process is very expensive. Louis [22]. developed Wcopyfind in which the process of comparison in this respect applied according to the units of phrase in the document, where the phrase structure contains six or more words The similarity is calculated by using count number of matched words from matching phrase over the total number of words in the same document. Heintze[15], Broder [2], Monostori [30] proposed a fingerprints methods to find the string matching and plagiarism detection based on common fingerprints proportion. These methods get good results but it fails when the plagiarized part is modified by rewording or changing some words of suspected parts. Ahmed H [3] introduced plagiarism detection using graph based representation. This method represented as an idea only without experiments or results, so the evaluation of this method is difficult without results and experiments. Pablo[31] proposed a system based on LempelZiv distance, which is applied to extract structural information from texts. The method seeks for the outliers in the vector of distances among each text fragments. Thomas [36] introduced a method based on standard information retrieval techniques by selecting an efficient data structures for the detailed analysis between the original and suspected document. Daniel [13] proposed a textual similarity method to capture plagiarism. This method ignored the semantic similarity between the original and suspected document. Through the related works, we noted the majority of plagiarism detection systems focused on the lexical structure and ignored the linguistic detection. On the other hand our proposed method focused on both lexical and linguistic aspects.

2.2. GRAPH-BASED REPRESENTATION

Nowadays, many of methods that are used for document representation rely on bag of word model (BOW) commonly known as a vector space model (VSM). The documents are represented as a linear vectors and the co-occurrence of the words in text document corpus. Many semantic relations among concepts and significant information are lost when a vector space model is used. If the document is long, it is very difficult to represent it as a vector model due to the large dimensionality. On the other hand, Latent Semantic Indexing (LSI) is an additional common method that focuses on transforming the source document vector to reduce the dimensional space using correction analysis structure of the terms in text document collection. Web document representation has been especially designed by Schenker [1]. The main benefit of graph-based method is that it allows keeping the structural information inherent to the source document. The methodology of graph-based representation contains definitions of graph based, sub graph and graph isomorphism. Based on [8] graph G is a 4-tuple G=(V, E, α, β) where V is a set of nodes (vertices), E is a set of edges connecting with nodes $E \subseteq V \times V$, α is a function Labeling the nodes $\alpha : V \rightarrow \sum v$, and $\beta : V \times V \rightarrow \sum e$ is a function Labeling the edges ,the ($\sum v$ and $\sum e$ being the sets of labels that can appear on the nodes and edges). In brief, we refer to G as G= (V, E) by "omitting the Labeling functions". A graph G1=(V1,E1,α1,β1) is a sub graph of a graph G2=(V2,E2,α2,β2), denoted $G1 \subseteq G2$, if $V1 \subseteq V2$, $E1 \subseteq E2 \cap (V1 \times V1)$, $\alpha1(x) = \alpha2(x) \ \forall x \in V1$ and $\beta1(x, y) = \beta2(x, y) \ \forall (x, y) \in E1$. Graph based representation proposed by Schenker [1] is based on the adjacency of terms in an HTML document. Under the standard method [1], each unique term (word) appearing in the document, except for stop words such as "the", "of", and "and" which convey little information, becomes a vertex in the graph representing that document. Each node is labelled with the term it represents. A single vertex for each word is created even if a word appears more than once in the text to build the terms' graph

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

in the sentence. Graph representation is an initial stage for text mining. It concentrates on how to represent text document as graph. The graph based representation relies on the processing of the text level. Zhang [38] divided the graph representation into three levels; document level, sentence level and term level. The representation of these levels as graph defined the graph node and graph edge. Since the node can hold the document or sentence or term and the edge is a weight between these levels. Document level looks at the multi documents in the graph. Here each document in the corpus or web is represented as a node and each relationship or link between two documents is demonstrated as an edge.

Many algorithms were employed for the document level graph-based such as page rank [32] and Hyperlink Induced Topic Search (HITS) introduced by [17]. The main task for the document level graph is how to represent the document in the graph to link the information on each other. These documents have a relationship among them through the information similarity between them. Second level works on the sentences level representation. In this level of graph representation, each node in the graph is represented as a sentence and each edge represents the relationship between two nodes or sentences. This relationship could be similar to that between sentences. One of the prominent fields that uses the sentence level is text summarization. [14] Used page ranking techniques with a sentence graph to select the high ranked sentence in text summarization. In term level, each term is represented by node and the relationship between the two terms could be co-occurrence for the terms. Some of studies concentrated the term level such as text clustering, text classification and summarization [26],[14], [40]. [26] Introduced web documents graph based representation by using the semantic and text location. He used frequent subgraph extraction algorithm to extract frequency of the document subgraphs [19]. He used the extracted subgraphs in documents classification. This method is similar to n-gram extraction technique inclusive of all kinds (one-gram, two-gram, and three-gram). According to term representation as node inside the graph, the important terms are considered in the representation for the graph. [40] Introduced the term graph representation in document clustering. He represented a bipartite graph based using documents clustering algorithm. The researchers and authors apply ontology techniques during the vector space representation construction using mapping of the terms graph for the documents to ontology and combine some concepts depending on the hierarchy of the concepts. Reciprocal support strategy is applied to recursively assign the documents and terms to their corresponding clusters. In their method, the documents are represented using co-occurrence concept couples, which was displayed for dimension lessening [40].

3. GRAPH-BASED METHOD FOR PLAGIARISM DETECTION

Plagiarism Detection using Graph-based Representation aims to detect the similarity between two sentences and possible semantic similarity between them. In this section we discuss the idea of proposed method. We first propose pre-processing cross suspected documents and original documents which was done by using text segmentation into sentences, stop words removal and stemming process. Then we represent each document as graph structure. The graph consists of nodes and edges. Each sentence is represented in one node. The relationship between the nodes is represented by the edge. The value of this edge is equal the overlapping between the concepts of two nodes. The overlapping between the sentences nodes calculated according to *Jaccard coefficient* which can be defined in the following equation:

$$Overlap(S1,S2) = \frac{(|CS1i| \cap |CS2j|)}{(|CS1i| \cup |CS2j|)} \quad (1)$$

Where n is instead of = Concepts of Document2; CSi = Concepts of Sentences1; CSj = Concepts of Sentences2.

The produced text graph-based representation does not represent the content of a text document as a graph only, but also captures the underlying semantic meaning in terms of the relationships among its concepts. All sentence nodes are connected to a unique node known as" Topic Signature ". The topic signature node is formed by extracting the concepts of each sentence terms. By using the WordNet, all hyperonym and synonymy will be extracted and participated as concepts of the terms. Grouping of the concepts of the sentence in one node reflect content of these node. The Advantage of Topic Signature node is that it quickly guides us to capture the suspected parts from the documents. Figure 1 shows the graph-based representation for the text.

In this figure, all the terms of each sentence after pre-processing step are collected in one node and then the concepts for these terms are extracted and grouped in such node. All the sentences nodes are connected with a Topic Signature node. Topic

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Signature node have index record for each sentence node to determine each concept belong to each sentence. The plagiarized parts are detected using Topic Signature in comparison process. The comparison process is conducted base on similarity calculation between the concepts inside the suspected Topic signature node and original topic signature node

The similarity between the original document and suspected document calculated by the following equations:

$$Sim(D1,D2) \sum_{i=1}^{n} s1i \sum_{j=1}^{m} s2j \quad ----- (2)$$

$$\text{Similarity between } (S1,S2) = \text{Overlap}(S1,S2) = \frac{(|CS1| \cap |CS2|)}{(|CS1| \cup |CS2|)} \quad ----(3)$$

$$Sim(D1,D2) \sum_{\substack{i=1,n \\ j=1,m}} (S1i,S2j) \quad ----- (4)$$

Where m= Concepts of document1; n= Concepts of Document2; CSi= Concepts of Sentences in document1; CSj= Concepts of Sentences in document2; D1= Original document; D2= Suspected document.

The weighting between the topic signature and the sentences calculated by the following equation:

$$\text{Overlap}(T.S,Si) = \frac{(|CT.S| \cap |Si|)}{(|T.S|)} \quad ----- (5)$$

Where T.S= Concepts of Topic Signature; Si = Concepts of Sentence i.

The benefit of the weighting between the Topic Signature and each sentence node is to determine the important node that can hold a large number of concepts. By this weight we can select just the important nodes because sometimes we find very large documents, so the graph will be bigger and the comparison will take a long time due to the huge of the concepts that are extracted from the documents terms. Figure 2 illustrates a comparison between the original document and suspected document.

4. METHODOLOGY DESIGN AND PLAGIARISM DETECTION

In this section we will discuss the methodology that followed to detect the plagiarism based on our text representation method. The following steps guide to capture the plagiarized part among suspected documents.

4.1. DATA PRE-PROCESSING

Pre-processing is one of the key steps for good results in dealing with problems in a natural language processing (NLP). Technology of stop words removal for deleting meaningless words will be used. Stemming algorithm is also applied to remove the affixes (prefixes and suffixes) in a word in order to generate its root word. In this aspect, this step extracts the significant words from the text and ignores the remaining words. This may adversely affect the similarity between documents.

4.1.1 REMOVING STOP WORDS

In information retrieval, stop words are the words that frequently occur in documents. These words do not give any hint values or meanings to the content of their documents such as (the, a, and,), hence they are eliminated from the set of index terms [33]. Salton and McGill [29] reported that such words comprise around 40 to 50% of a collection of documents text words. Eliminating the stop words in automatic indexing will speed the system processing, saves a huge amount of space in index, and does not damage the retrieval effectiveness [37]. There are various approaches used for determination of such stop words list having the same aim which is to find those of no content values. Nowadays, there are several English stop words list that are commonly used to assist in information retrieval. This study has been carried out to extract all the stop words in the documents.

4.1.2 WORDS STEMMING

One of the problems involved in information retrieval is variation in word forms [21]. The most common types of variation are spelling errors, alternative spelling, multi-word construction, transliteration, affixes, and abbreviations. These variations in words form lead to the efficiency issue in the matching algorithm during information retrieval process. One way to overcome such problem is to use stemming. Stemming is a process to remove the affixes (prefixes and suffixes) in a word in order to generate its root word for example (learning) will be (learn) and learned will be learn also. Using root word in pattern matching provides a much better effectiveness in information retrieval. Nowadays, there are many stemmers available for the English language and are quite complete and thorough. For example, Nice Stemmer, Text Stemmer and Porter Stemmer are the well-known English stemmers which have been commonly used.

4.2. CONCEPTS OF THE TERMS

Concept identification is common to applications such as ontology learning, glossary

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

extraction and keyword extraction. These applications have different definitions for concept, hence different methods. Previous methods start from the idea that concepts can be found as words or phrases contained in sentences. They are then divided into smaller phrases in one of two ways, Grammatical or Syntactical information. The former can be found in ontology learning, glossary extraction and information retrieval systems. Using a shallow grammar parser, an entire sentence is phrased into a grammatical tree which classifies sub-phrases as noun or verb phrases. Noun phrases are selected as concepts. The syntactical information division of sentences uses punctuation or conjunctions to separate phrases within a sentence, all these phrases are concepts. In this study, the concept extraction is carried out using a WordNet thesaurus by extracting the hyperonym and synonymy for the terms.

4.3. SIMILARITY DETECTION AND GRAPH MATCHING

This step has been conducted by breaking down the suspected and original documents into their constituent sentences. Pre-processing for each document such as a segmenting of each sentence into separated terms is required. Stop words removal and concepts extraction for each term within a sentence follows. This method represented the sentences in the form of nodes. Each node contains one sentence from the document. Section (3) motioned how to represent the terms of each sentence as graph. However, each node in the subgraph contains one term for the concepts grouped in the topic signature. By this grouping the similarity between the words and concepts had been detected, because sometime people attempt to hide their activity and change words by replacing synonyms thereby modifying a structure of sentences. Some plagiarism detection systems fail in matching overlap with an original document if the replacing done. The proposed method uses a concept extraction to avoid this problem using hyperonym and synonyms of words from the WordNet Thesaurus dataset. Tests have been carried out using PAN-PC-09 standardization of plagiarism detection corpus. The major benefit of the proposed method is that the graph acts as quick guide to the related suspected nodes of the sentences. We have found by experiment that our proposed method achieves a good performance compared to the others techniques such as semantic based method and Longest Common Subsequence (LCS). Figure 3 illustrates the structure phase of methodology.

5. EXPERIMENTAL DESIGN AND RESULTS EVALUATION

The experiments play a very important role in this study. It looked at the amount of detected plagiarized sentences from the original documents. The following steps explain how the proposed method works. In the beginning, the suspected and original documents will break down into their constituent sentences. Pre-processing for each document such as a segmenting of each sentence into separated terms is required. Stop words removal and concepts extraction for each term within a sentence follows. Then the sentences will be represented in the form of nodes. Each node contains one sentence from the document. All nodes connected by Topic signature node. The comparison between the suspected and original document calculated based on the similarity between the Topic Signature of suspected and original document.

The experiment looks at the amount of detected plagiarized sentences from the original documents. The major benefit of the proposed method is that the graph acts as quick guide to the related suspected nodes of the sentences. The experimental results on PAN-PC-09 dataset show that our method remarkably outperforms the modern methods for plagiarism detection in term of Recall, Precision and F-measure. We provide 3 general testing parameters that commonly used in plagiarism detection area nowadays in testing phase.

$$Recall\ (R) = \frac{number\ of\ detected\ plagiarized\ sentences}{total\ number\ of\ sentencee}$$

eq. (6)

$$Precision\ (P) = \frac{number\ of\ corrected\ detected\ plsgiarized\ sentences}{number\ of\ detected\ plagiarized\ sentences}$$

eq. (7)

$$F\text{-}measure = \frac{2\ x\ R\ x\ P}{R+P}$$

eq. (8)

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Tables 1 illustrate a result cross the 100 documents.

Documents	Number of sentences	Plagiarized Sentences	Non-plagiarized Sentences	Detected Sentences	Non-detected Sentences
Doc 1	45	14	31	45	0
Doc2	157	139	18	155	2
Doc3	29	25	4	26	3
Doc4	42	26	16	40	2
Doc5	62	33	29	65	2
Doc6	44	53	24	72	5
Doc7	49	46	3	49	0
Doc8	52	35	17	43	9
Doc9	43	26	17	42	1
Doc10	44	15	29	44	0
Average	60	41.2	18.8	58.1	2.4

Table 1 Sample of results cross the 100 documents

In table 1, we first calculate the number of sentences inside the documents, and then we check how many sentences exactly plagiarized from the original documents. Our proposed method is detecting the number of sentences similar to the original sentences. This detection includes sentences exactly plagiarized and some of them just similar but actually not plagiarized. Also we noted in the table there are some sentences plagiarized but our proposed method can not detected because these sentences changed totally such as style and structure, where our proposed method focused on copy paste and paraphrasing types. We used this table as an input to calculate the Recall, Precision and F-measure for the proposed method in the next section.

6. RESULTS EVALUATION

The current experiment was performed on 100 documents. Each suspected document is plagiarized from one or more different original documents. The experiment searched for plagiarized part of the suspected documents from the original documents. All the documents were collected from PAN-PC-09 standardization of plagiarism detection corpus. Those suspected documents are plagiarized with different plagiarism ways such as simple copy and paste, change some terms to its corresponding synonyms, and modify the structure of the sentences (paraphrasing).

These parameters used to evaluate the performance of our proposed method.

Our proposed method was evaluated and compared with some of the modern techniques in plagiarism detection. The next table 3 shows the comparison between our proposed method and other techniques.

Evaluation measure	Longest Common Subsequence (LCS)	Semantic Technique	Graph based model
Recall	0.6111	0.7222	0.9615
precision	0.6667	0.7778	0.6886
F-measure	0.6378	0.7490	0.822

Table 2 Comparison between LCS, Semantic Technique and Graph-Based Model

Table 2 demonstrates the comparison between Graph-based Model Longest Common Subsequence (LCS) [4] and Semantic-based similarity [11]. We found our proposed method achieved better results in recall, precision and f-measure than LCS and better results in recall and f-measure compared with Semantic technique. Figure 4 shows the comparison results among some of the plagiarism detection techniques.

Figure 4 describes the results and evaluation when we compare our proposed method with some of plagiarism detection techniques such as LCS and Semantic-based technique. We noted that our proposed method significantly outperforms the LCS and Semantic-based methods in term of Recall and F-measure.

The two curves in figures 4 and 5 illustrate the precision, recall and f-measure for the suspected documents when the comparison done which the suspected parts was captured. Figure (5) illustrates the comparison between the proposed method with LCS and Semantic-Based techniques. This comparison proved that the introduced method is the best through the experiment in Recall and F-measure and better than LCS in the all of the evaluation measurement. Table (3) illustrates the results evaluation using a Recall, Precision and F-measure as an evaluation measures that applied to prove the performance of the proposed method.

7. CONCLUSION

In this paper the problem of plagiarism detection has been considered one of the most publicized forms of text reuse around us today. In

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

particular, it has been shown in this paper how plagiarism detection is handled using graph-based representation. The graph based model does not represent the content of a text document as a graph only, but also captures the underlying semantic meaning in terms of the relationships among its concepts. The graph was produced by grouping each sentence terms in one node. The resulting nodes are then connected to each other based on the order of sentences within the document. All nodes in the graph are connected to a top level node known as" Topic Signature ". The topic signature node is formed by extracting the concepts of each sentence terms using WordNet thesaurus hyperonym and grouping them in that node. Tests have been carried out using PAN2009 standardization of plagiarism detection corpus. The major benefit of the proposed methods is that the graphs act as quick guide to the related suspected nodes of the sentences. The proposed method has been found to achieve a better performance in term of effectiveness by the experiment and comparing with LCS and Semantic-based techniques.

REFRENCES:

[1] A. Schenker, H. Bunke, M. Last, and A. Kandel, "Graph- Theoretic Techniques for Web Content Mining", Series in Machine Perception and Artificial Intelligence, 62, World Scientific, 2005.

[2] Z. Broder. On the resemblance and containment of documents. In Proceedings of Compression and Complexity of Sequences. IEEE Computer Society pages 21-29, 1997.

[3] Ahmed H. Osman, Salim N. , Binwahlan, M. S. Salem.Plagiarism Detection Using Graph-based Representation. Journal of Computing, Volume2 , Issue 4, ISSN 2151-9617, 2010.

[4] Alzahrani, S., Salim, N., Kok Kent C., Binwahlan, M. S., and Suanmali, L. The Development of Cross-Language Plagiarism Detection Tool Utilising Fuzzy Swarm-Based Summarisation. In International Conference on Intelligent Systems Design and Applications ISDA 2010.

[5] Angélil-Carter, S. (2000), Stolen Language - plagiarism in writing, Real Language Series, Pearson Education Limited.

[6] Antonio, S., Hong Va, L., & Rynson, W. H. L. CHECK: a document plagiarism detection system. Paper presented at the Proceedings of the 1997 ACM symposium on applied computing, 1997.

[7] Brin, S., J. Davis, and H. Garcia-Molina, Copy Detection Mechanisms for Digital Documents. ACM. p. 398-409, 1995.

[8] Bunke, H.; Jiang, X.; and Kandel, A. On the Minimum Common Supergraph of Two Graphs. Computing 65(1): 13-25.2000

[9] Brin, S., J. Davis, and H. Garcia-Molina, Copy Detection Mechanisms for Digital Documents. ACM. p. 398-409, 1995.

[10] Bunke, H.; Jiang, X.; and Kandel, A. On the Minimum Common Supergraph of Two Graphs. Computing 65(1): 13-25. 2000

[11] Chow Kok Kent, Naomie Salim, "Web Based Cross Language Plagiarism Detection," cimsim, pp.199-204, Second International Conference on Computational Intelligence, Modelling and Simulation, 2010.

[12] Clough. P. Plagiarism in Natural and Programming Languages: an Overview of Current Tools and Technologies. Research Memoranda: CS-00-05, Department of Computer Science. University of Sheffield, UK, 2000.

[13] Daniel Micol, Óscar Ferrández, Fernando Llopis, Rafael Muñoz. in Proceedings of the Uncovering Plagiarism, Authorship, and Social Software Misuse PAN 2010 Workshop 2010.

[14] Erkan, G., and Radev, D.R. "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization." J. Artif. Intell. Res. (JAIR) 22: 457–479. 2004

[15] Heintze, N. Scalable document fingerprinting. Paper presented at the Second USENIX Workshop on Electronic Commerce, 1996.

[16] Kang, N., Gelbukh, A., Han, S.-Y. PPChecker: Plagiarism pattern checker in document copy detection. In: Sojka, P., Kopeˇcek, I., Pala, Keds. TSD. LNCS LNAI, vol. 4188, pp. 661–667. Springer, Heidelberg, 2006.

[17] Kleinberg, J.. "Authoritative sources in a hyperlinked environment." In Proc. Ninth Ann. ACM-SIAM Symp. Discrete Algorithms, pages 668–677, ACM Press, New York. 1998.

[18] Krisztian Monostori, Arkady Zaslavsky, Heinz Schmidt "Document Overlap Detection System for Distributed Digital Libraries", Proceedings of the fifth ACM conference on Digital libraries, pp. 226 – 227, 2000.

[19] Kuramochi M. and Karypis, G. "An Efficient Alogrithm for discovering frequent Subgraphs."

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

IEEE Trancations on Knowedge and Data Engineering 16, 9 (Sep 2004).

[20] LaFollette, M. C. (1992), Stealing into Print: Fraud, Plagiarism, and Misconduct in Scientific Publishing. Berkeley: University of Calfornia Press.

[21] Lennon, M., Pierce, D.S., Tarry, B.D. and Willett, P,"An evaluation of some conation algorithms for information retrieval", Journal of Information Science, Vol. 3 No. 4, pp. 177-83, 1981.

[22] Louis Bloomfield, http://plagiarism.phys.virginia.edu. The Plagiarism Resource Site Charlottesville, Virginia.

[23] Lyon, C., Malcolm, J. A., & Dickerson, R. G. Detecting short passages of similar text in large document collections. Paper presented at the Conference on Empirical Methods in Natural Language Processing, 2001.

[24] Lyon, C., Barrett, R., Malcolm, J. A Theoretical Basis to the Automated Detection of Copying Between Texts, and its Practical Implementation in the Ferret Plagiarism and Collusion Detector. In: Plagiarism: Prevention, Practice and Policies Conference, Newcastle, UK. 2004.

[25] Manuel, Z., Marco, F., Massimo, M., & Alessandro, P. Plagiarism Detection through Multilevel Text Comparison. Paper presented at the Second International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution. 2006.

[26] Markov, A., Last, M., and Kandel, A. "Model-based classification of web documents represented by Graphs." Proceedings of WebKDD 2006 workshop on knowledge discovery. 2006.

[27] Martin, B. Plagiarism: a misplaced emphasis, Journal of Information Ethics, Vol. 3(2), 36-47. 1994.

[28] Meyer zu Eissen, S., Stein, B. Intrinsic plagiarism detection. In: Lalmas, M., Mac-Farlane, A., R¨uger, S.M., Tombros, A., Tsikrika, T., Yavlinsky, Aeds. ECIR 2006. LNCS, vol. 3936, pp. 565–569. Springer, Heidelberg, 2006.

[29] Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth, Using predicate-argument structures for information extraction. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, 2003.

[30] Monostori, K., A. Zaslavsky, et al. Document overlap detection system for distributed digital libraries. In Proceedings of the ACM Digital Libraries, 2000.

[31] Pablo Suárez, José Carlos González, Julio Villena-Román, in Proceedings of the Uncovering Plagiarism Authorship and Social Software Misuse PAN 2010 Workshop 2010.

[32] Page, L., Brin, S., Motwani, R., and Winograd, T. "The PageRank citation ranking. 1998.

[33] Rijsbergen, C.J. Van,. A New Theoretical Framework for Information Retrieval. Department of Computing Scirnce University of Glasgow, 1979.

[34] Shivakumar, , N. and H. Garcia-Molina, SCAM: A Copy Detection Mechanism for Digital Documents. D-Lib Magazine, 1995.

[35] Si, A., Leong, Lau, H., & Rynson W. CHECK: A Document Plagiarism Detection System. ACM Symposium for Applied Computing, pp.70-77, Feb. 1997.

[36] Thomas Gottron . in Proceedings of the Uncovering Plagiarism Authorship and Social Software Misuse PAN 2010 Workshop 2010.

[37] W.B. Frakes and R. Baeza-Yates, Information Retrieval: Data Structures and Algorithm. Department of Computer Science, University of Chile. 1992.

[38] Xiaodan Zhang. Exploiting External/Domain Knowledge to Enhance Traditional Text Mining Using Graph-based Methods. PhD. Drexel University. 2009.

[39] Yerra, R., & Ng, Y.-K. A Sentence-Based Copy Detection Approach for Web Documents. In Fuzzy Systems and Knowledge Discovery pp. 557-570, 2005.

[40] Yoo I., Hu X., and Song I-Y. "Integration of Semantic-based Bipartite Graph Representation and Mutual Refinement Strategy. 2006.

Journal of Theoretical and Applied Information Technology
31st October 2011. Vol. 32 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

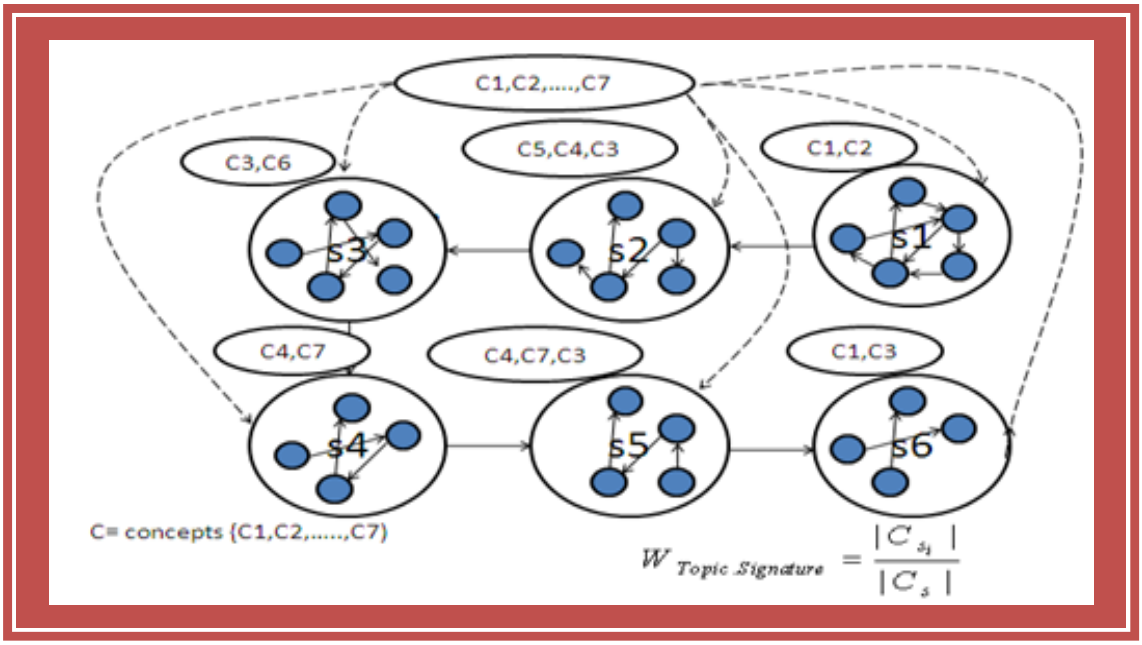

FIGURE 1

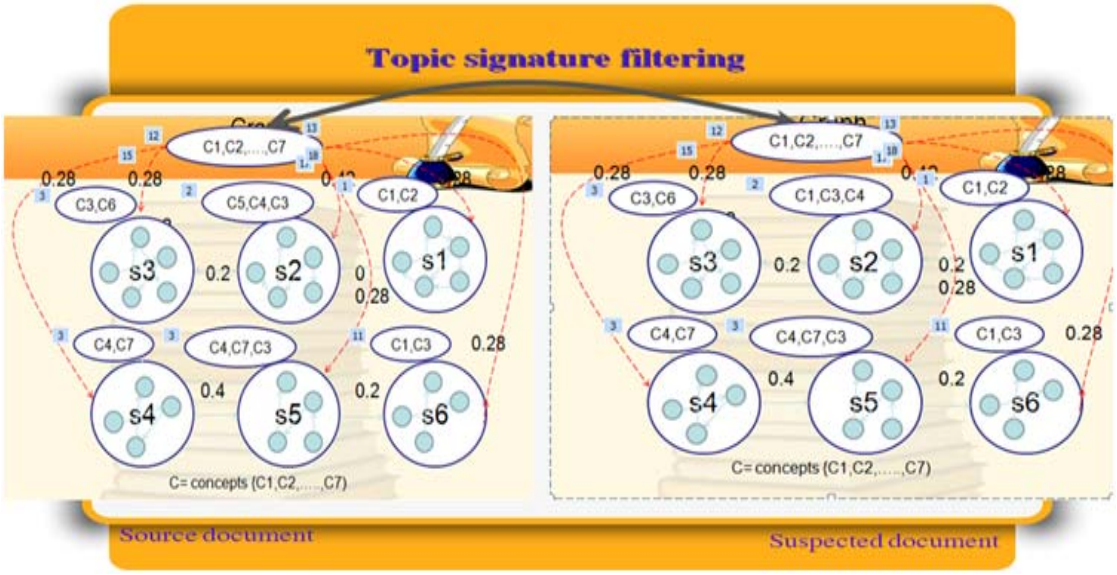

FIGURE 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

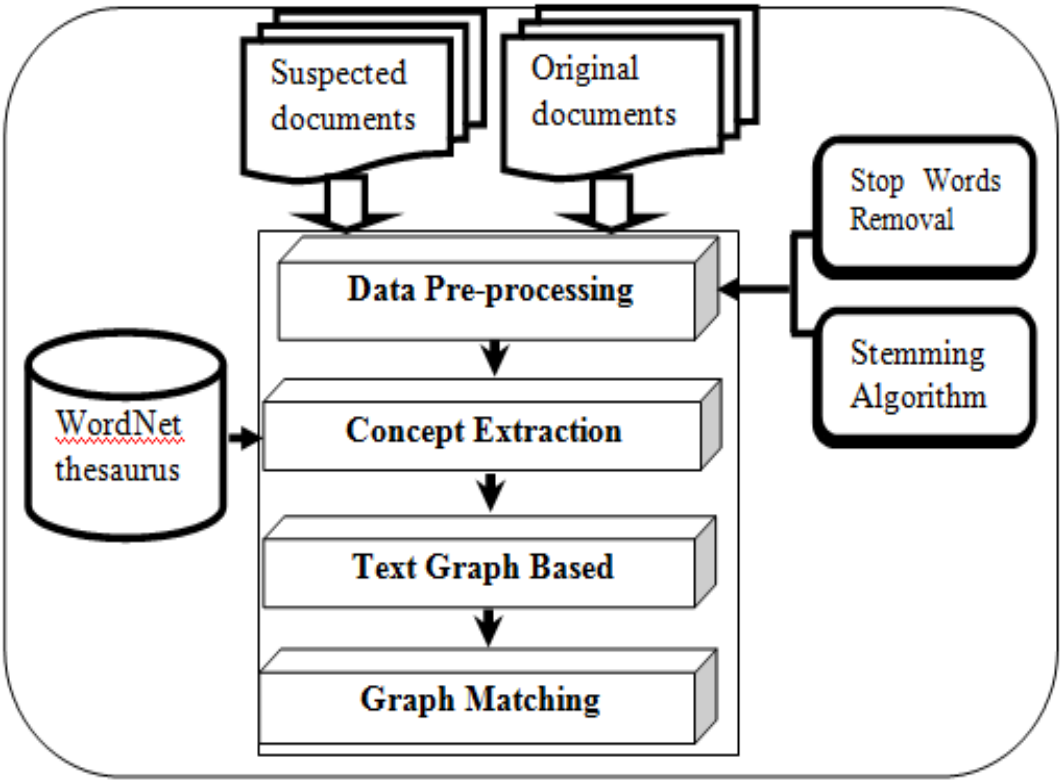

Figure 3

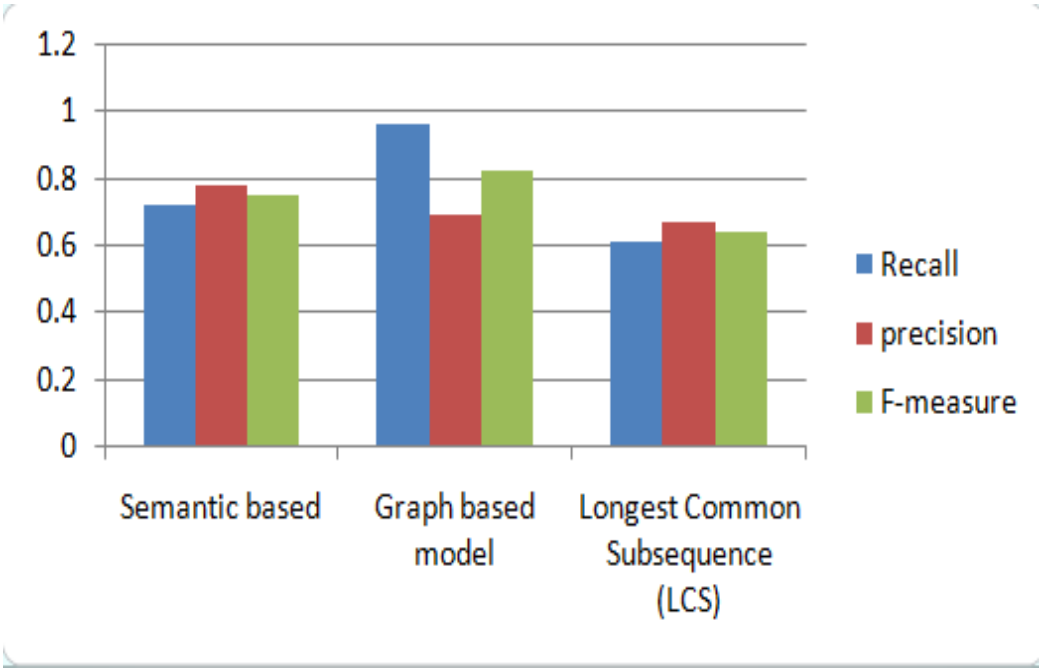

FIGURE 4