

TEXTUAL AND STRUCTURAL APPROACHES TO DETECTING FIGURE PLAGIARISM IN SCIENTIFIC PUBLICATIONS

IDRIS RABIU, NAOMIE SALIM

Faculty of Computing University Technology Malaysia
Idrisrabiu76@yahoo.com

ABSTRACT

The figures play important role in disseminating important ideas and findings which enable the readers to understand the details of the work. The part of figures in understanding the details of the documents increase more use of them, which have led to a serious problem of taking other peoples' figures without giving credit to the source. Although significant efforts have been made in developing methods for estimating pairwise diagram figure similarity, there are little attentions found in the research community to detect any of the instances of figure plagiarism such as manipulating figures by changing the structure of the figure, inserting, deleting and substituting the components or when the text content is manipulated. To address this gap, this project compares the effectiveness of the textual and structural representations of techniques to support the figure plagiarism detection. In addition to these two representations, the textual comparison method is designed to match the figure contents based on a word-gram representation using the Jaccard similarity measure, while the structural comparison method is designed to compare the text within the components as well as the relationship between the components of the figures using graph edit distance measure. These techniques are experimentally evaluated across the seven instances of figure plagiarism, in terms of their similarity values and the precision and recall metrics. The experimental results show that the structural representation of figures slightly outperformed the textual representation in detecting all the instances of the figure plagiarism.

Keywords: *Plagiarism., Figures, Images, Pairwise Diagram, Jaccard Similarity Measure*

1 INTRODUCTION

The problem of plagiarism has recently increased because of easy access to the web, large databases, and telecommunications in general. It becomes very easy for people to browse and access any information of their interest through the web pages. This however turned plagiarism into a serious problem for students, publishers and researchers in educational institution. There are many definitions of what constitutes plagiarism. According to [1] plagiarism occurs when the word, ideas, diagrams, designs, photographs, maps, graphs, verbal communication of information, derived equations, computer programs, illustrations, tables and primary data of another is passed off as your own, unless the source is acknowledged and properly documented. This definition applies regardless of which medium the source material is published, including any material copied from soft copy publications (i.e. Internet, email attachments, e-journals, etc.) and hard copy publications (such as textbooks, these, journals

etc.). Although the furthermost common type of plagiarism is the document text in which the plagiarized documents is made by copy-pasting all or some parts of the original document, plagiarism can also be found in many aspects which includes the concepts and figure plagiarisms. Several research efforts have proposed different methods for detecting different kinds of plagiarism, which mostly are centered on text analysis such as string matching, fingerprinting or style comparison of the documents. Code clones detection methods have also been in existence since 1970s according the research studies, to detect programming code plagiarism detections in Pascal and C languages [2]. Figure plagiarism detection, in particular, flowcharts and framework diagrams detection is largely unexplored. Figures are essential parts of scientific paper that are often used to present complex results in a readable way. The role of figures in disseminating important ideas and findings which enable the readers to understand the details of the documents guarantees more use of them, which in the other way increase the rate

of figure plagiarism. The figure plagiarism can be considered as the act of copying the information about figures from another person's work without citing the source. Detection of figure plagiarism however, is a complex problem in that, as opposed to document plagiarism detection which is usually done by extracting and comparing the texts, it is not only limited to text comparison but also the extraction and comparison of all the noticeable visual relationships that exist between all the components of the figures. Therefore, the major challenge in the figure plagiarism detection is how to extract and represent the information from figures in order to detect the sections of plagiarism. In this paper we consider two ways to represent the figure for the purpose of plagiarism detection: textual representation and structural representation. Textual representation allows us to compare the figures by comparing the vocabulary of terms in each component; whereas, the structural representation allows us to compare the figures using the vocabulary of terms and keeping the track of relationships that exist between the components. For each component in each figure, several searches are done, trying to disclose all the possible forms of a plagiarism. The textual similarity detections are obtained with simple n-gram technique and Jaccard similarity measures; and graph edit distance measures for structural similarity detection. To evaluate the effectiveness of the proposed detection system, precision and recall are implemented.

2. DIFFERENT CATEGORIES OF FIGURE IN SCIENTIFIC PAPERS

The goal of this paper is to extract the useful information from figures and to represent the extracted features in a way they can be easily searched. To facilitate further analysis of figures, we first define the semantic types of figures in scientific documents based on their features and the functionalities. The figures in scientific papers are categorized as photographs and non-photographs. Figure 1 shows a tree structure of major categories of figures.

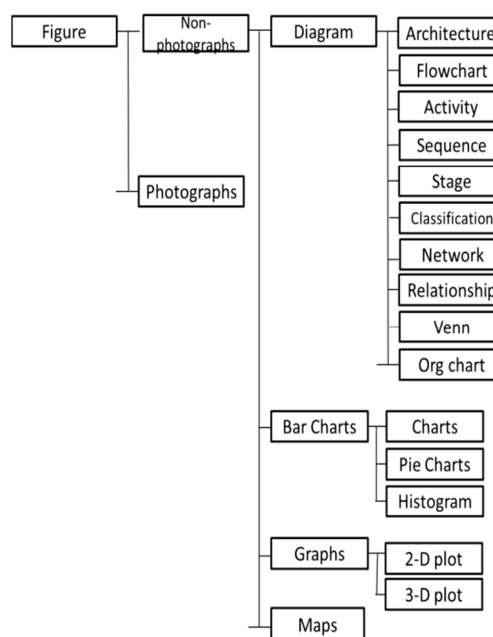


Figure1: Categories of Figures in Scientific Documents

2.1 Photograph:

A photograph or photo is an image created by light falling on a light-sensitive surface, usually photographic film or an electronic imager such as a charge-coupled device (CCD) or an active pixel sensor (CMOS) chip. Most photographs are created using a camera, which uses a lens to focus the scene's visible wavelengths of light into a reproduction of what the human eye would see. The process and practice of creating photographs is called photography. Figure 2 shows some samples of photographs. This category however is not the focus of this research and it is therefore not detailed here.

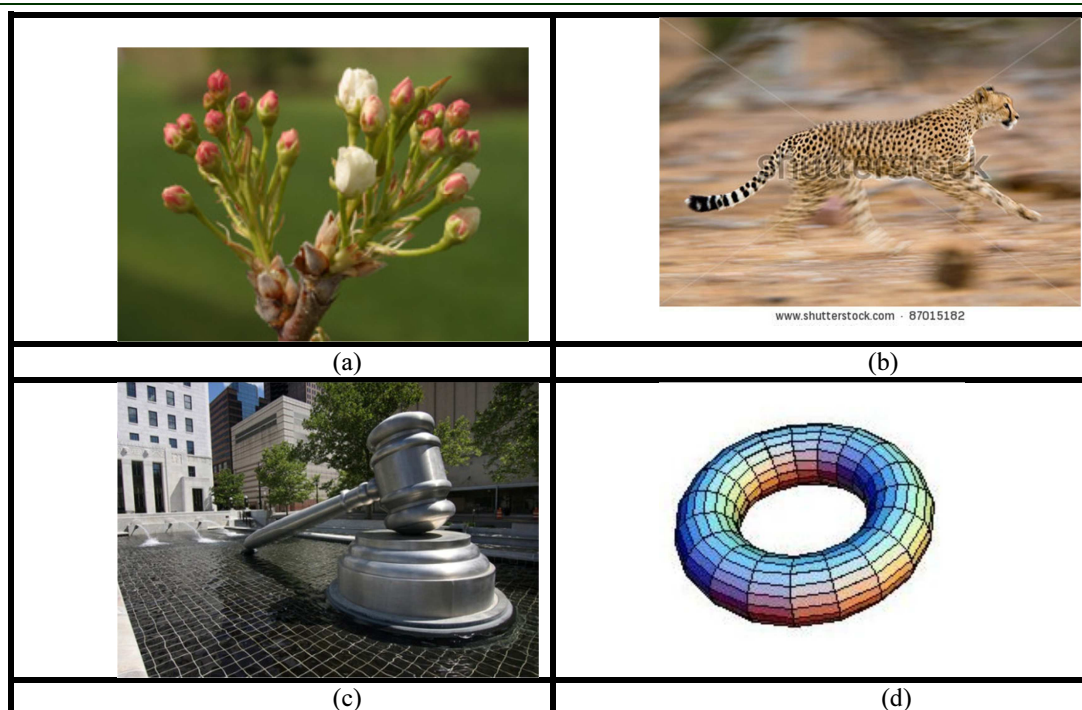


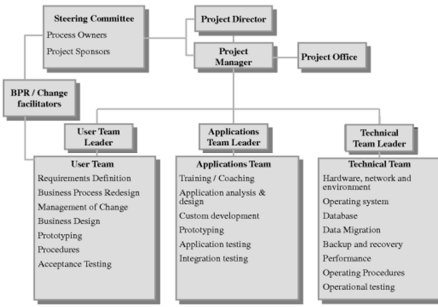
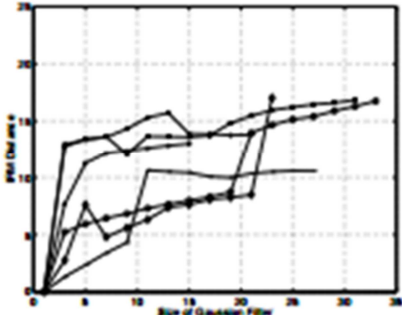
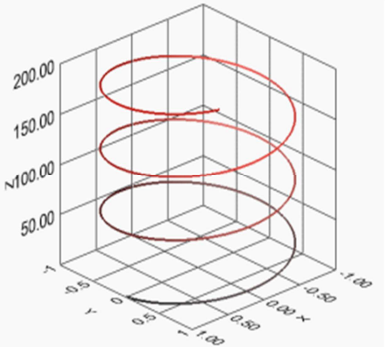
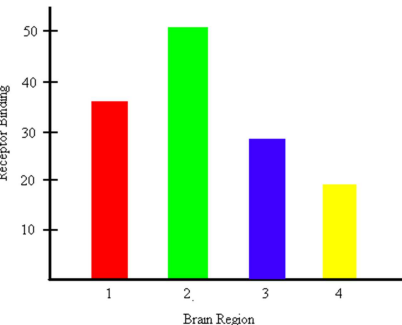
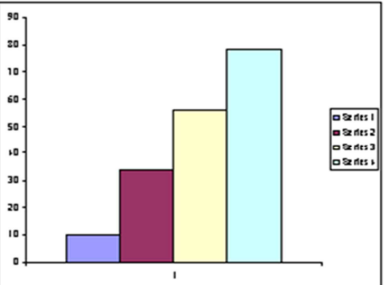

Figure 2: Some figures categorized as “photograph”.

2.2 Non photograph figures

Non photograph figures are the second category of figures which can further be categorized as diagrams, graphs, charts and maps. The diagrams show the arrangement and relational dependencies among a series of component illustrating the key ideas from scientific documents. The components in the diagram are usually represented by closed contours such as: rectangles, oval, diamonds, etc[3]. Detection of this category is the focus of this research. Graphs and charts are used to represent numerical information in a pictorial or illustrative form, allowing better understanding of the data. The notable examples of charts are bar charts, histogram charts, and pie charts which are used to illustrate the relative proportion of each category of multiple variable data with all others. Graphs are different from Charts as they are built by plotting the values of a function along an axis

that represent some possible values taken by a variable. Common representations of graphs are either functions plotted over two orthogonal axes i.e. 2-D plots or a projected 3-D plot over three axes. Maps are diagrammatic representations of objects, data, and regions showing the distribution or special arrangement of an area. Table 1 shows the pictures of the non-photographic figure categories that are discussed above.

Table 1: Sample of Non-Photograph Categories of figures

	
<p>A sample of Diagram figure</p> 	<p>2-D plot</p> 
<p>3-D plot</p> 	<p>A sample of Bar chart</p> 
<p>Histogram chart</p>	<p>Samples of Map figures</p>

3. Related Works

In the literatures, many techniques and tools have been developed to address two different types of document plagiarisms: programming source code documents [4-9] and natural language documents [10-18]. In the same vein and due to the fact that the description of figure could be done in natural languages, information storage and retrieval theories were applied to storing and retrieving the diagrams figures such as software reuse and evolution [19], model management [20], and collaborative design and development [21]. Detecting figure plagiarism is however a challenging task because different

types of figures exist such as graphs, charts, photographs, maps and diagrams which is the focus of this paper. Therefore different features can be extracted from different figures for the purpose of comparison depending on the target of the comparison. For example in the figure that contains a graph of a 2-D plot or 3-D plot, the number of curves and the data contained in each curve can be used for similarity search[3]. In chart similarity detections, shape detection is performed and the general shape descriptors are used to form the feature vectors. For example, [22] developed a prototype system for similarity matching between statistical charts (bar charts, pie charts and line charts) based on the special

features of each category of charts. Some of the common features that can be used for bar chart similarity includes: the number of bars, the width of each bar, and height of each bar, horizontal or vertical alignment of base, color, sequence and textual annotation. For simple pie charts, the features include the number of slices, angle of slices, area of each slice, the color of each slices and textual annotation. Last of all, the line charts, have the features such as number of lines, number of silent points on a line, position of each silent points, color of lines and textual annotation. In the figure that contain diagrams, the organization of diagrams, the texts and the number and the flow of components can be used as features of interest in the attempts to detect similarity in this category of figures[3]. In diagram similarity estimation literature, many efforts have been devoted to developing the algorithms and methods for detecting different types of diagram by focusing on text contents of diagrams such as use case diagrams, workflow diagrams, sequence diagrams and process models diagrams. For example, [23] proposed an automated technique for calculating the sequence diagram similarity with a supporting tool called "ScenAsst" for storing and retrieving use cases. ScenAsst transforms the content of the diagram into a graph. In retrieval process, the query diagram area also transformed into a graph and similarities between them and each graph in the collection is compared by using "SubDue" algorithm. [24] presented the methods for measuring the similarity between the business process model diagrams. These methods derived similarities based on three dimensions: syntax, linguistic and structural similarity measures.[25] studied the problem of workflow diagram similarity estimation and retrieval by introducing a structure-based approach using a weighted graph edit distance. [26] proposed the CBR to retrieve use case diagrams. In their approach the retrieval methods are designed to match the use case diagrams taking into account two dimensions: use case and actor dimension (i.e. Text based format) and the relationship dimension (i.e. Structure based format). The matching score and weight of each dimension are calculated based on the nearest neighbour matching and ranking to find the most similar diagrams. [27] evaluated and compared three classes of similarity measurement methods for business process model diagrams which are: label similarity (text similarity), structural similarity (text as well as the relationship

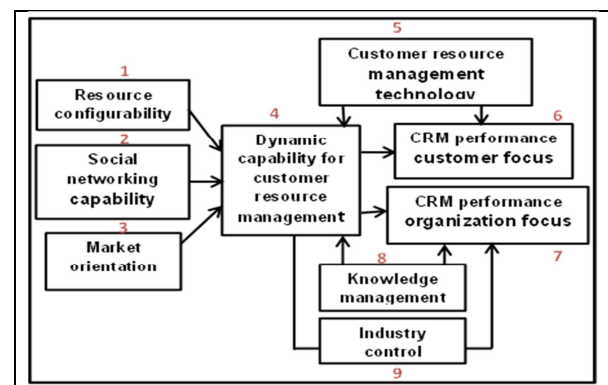
between the components) and behavioural similarity (the text as well as the causal relations captured in a process model). Based on their experiments the results show that the structural similarity achieves the best performance from the three representations.

4. Figure Representation

This section introduces the new paradigm for figure plagiarism detection and its supporting frameworks. It explains the proposed techniques for representing the figures in a way that support detection of the plagiarized portions in figures. The proposed method involves two approaches: text-based approach- usually represented as the plain texts extracted from the figures and structure-based approach- usually represented as the texts as well as the relationships that connect the components, which were inspired from [26].

4.1 Textual Representation

The textual representation of the figure is the concept of representing the figures as a text document based on the textual information that surrounds the figures. In many of the diagram similarity estimation research, the methods of representation for retrieving the similar diagrams were solely relying on the this approach because the description of each component of the figure was represented as text-based such as the works of [28]. Extensive research has been made in the work of [29] on the methods that find similarities in the use case diagrams figures both in the text-based and the relationship-based dimensions. The text in a figure can sometimes be a single word or phrase. For this reason the content of the figure for similarity detection are mostly represented based on a character-based, word-based or phrase-based. The figure 3 shows the example of figure and its text representation.



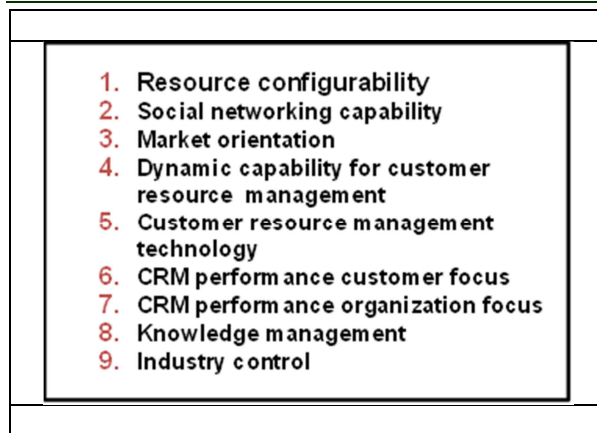


Figure 3: Example of figure and its text representation

4.2 Structural Representation

The structural representation is another approach for representing the figures which cover the entire structure of the figure in determining their similarity. It takes into account both the text within the components of the figure as well as the relationship between them. The Structure based approach has been used for a variety of language processing applications such as a parsed text similarity measure[30], diagram similarity estimation[24, 25, 27, 31-33], Schema element mapping [34] and plagiarism detection [35]. The notion of this metric is that the two matching figures are represented as labelled graphs, and the components as the nodes, and the link between them as the edges of the graphs. Figure 4 shows the examples of figures represented in graphical forms. The definition of figures in graphical form is given as follows:

Definition 1 (Diagram): A diagram G is a tuple $(N(G), E(G), \text{and } l)$ and Ω is a set of text labels, in which:

- $N(G) = \{N_i / N_i \in G\}$ is the set of nodes in G
- $E(G) = \{E_{ij} / N_i, N_j \in N(G) \text{ is the set of edges in } G\}$
- $l: (N \rightarrow \Omega)$ labels components with text.

Diagram G contains three types of data: nodes, edges and attributes. The nodes are referred to the components of the figures, edges are the relationship between the two components and attributes are the data associated with the nodes which in this case are strings of text.

Definition 2 (Relationship between diagram nodes): Let $G = \{N(G), E(G)\}$, where $N(G) = \{N_i / N_i \in G\}$ is the set of nodes in G . $E(G) = \{E_{ij} / N_i, N_j \in N(G)\}$ is the set of edges in G . For each node $n \in N$, the path $a \rightarrow b$ refers to the relationship that exist in a sequence of graph nodes $n_1, n_2, n_3, \dots, n_k \in N$ with $a = n_1$ and $b = n_k$ such that for all $i = 1, 2, 3, \dots, k$, the set of edges $(n_1, n_2), (n_2, n_3), (n_3, n_4), \dots, (n_{k-1}, n_k) \in E$ holds. For example, the relationships illustrated in figure 4 can be represented as: $d1 \rightarrow d2, d1 \rightarrow d3, d2 \rightarrow d4, d2 \rightarrow d5$, and $d2 \rightarrow d6$ for original figure, and $p1 \rightarrow p2, p1 \rightarrow p3, p2 \rightarrow p4$ and $p2 \rightarrow p5$ for plagiarized figure.

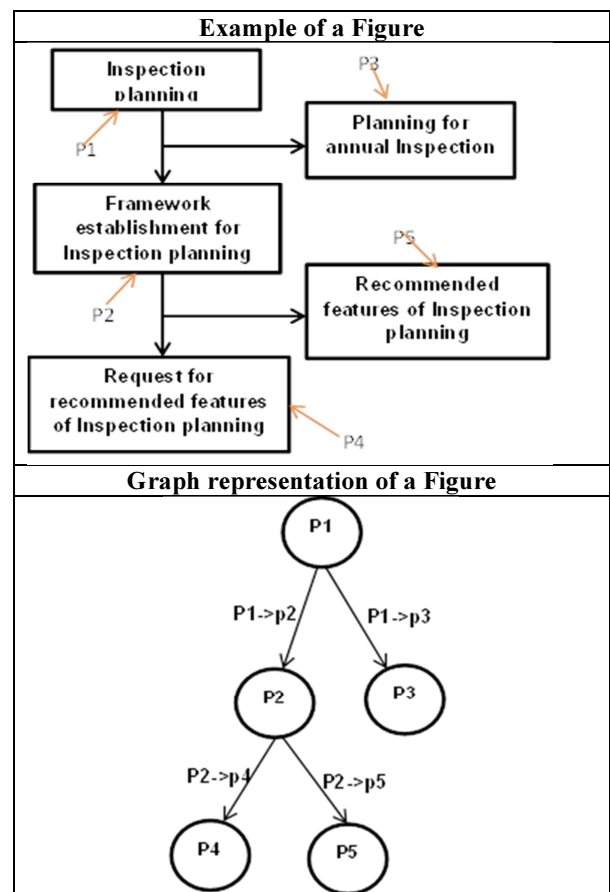


Figure 4: Example of a figure and its graph form representation

5. Figure Plagiarism Detection Methods

This section involves calculating the similarity between the terms in the each figure so as to determine the level of plagiarism in figures. Two methods will be exploited in determining the similarities. The first method is the textual

similarity matching, in which case, the texts will be compared on the basis of n-grams fingerprint method, using the Jaccard similarity measure. The second method is the structural similarity- using the graph edit distance measure which covers the both the text and relationship similarity matching. Figure 5 shows the summative diagram of the whole calculation process.

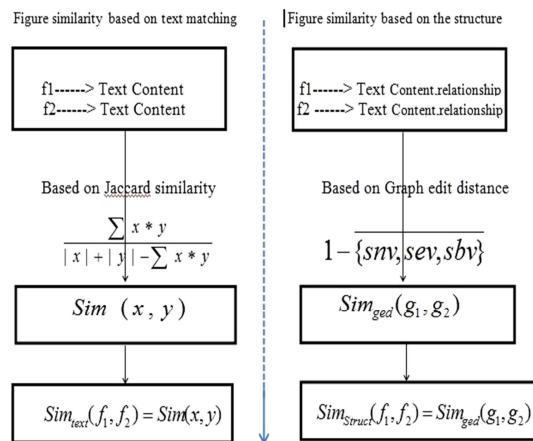


Figure 5: Summative Diagram for whole calculation process

5.1 Text Similarity Matching

Textual matching refers to the concept of matching between texts in order to display their similarity. However, the text in a figure can sometimes be a single word or phrase. For that reason, the methods of representing the content of the figure for similarity detection were mostly based on the character-based, word-based, or phrase-based. For example, [36] investigated two representations based on longest characters substrings (LCS) and a 2 word N-gram representation for measuring the syntactic similarity of the content of figures. To determine the similarity between the two figures on the textual similarity basis, we proposed n-gram representation of text documents contained in each figure. The n-gram based string matching technique is a representation in which the texts are decomposed into a successive set of words or string of characters, such as uni-gram which represent one word, bi-gram represent two words; three-grams represent three words etc. For example, the sentence “It is common for large and complex organization” can be sliced into 3-gram as {“It is common”, “is common for”, “common for large”, “for large and”, “large

and complex”, “and complex organization”}. Based on the n-gram selected, the similarity between the texts can be estimated using jaccard similarity measure defined as:

$$Sim(x, y) = \frac{|x \cap y|}{|x \cup y| - |x \cap y|} \dots \dots \dots (1)$$

Where $|x \cap y|$ is the number of words that are common in two sentences x and y , and $|x \cup y|$ are number of n-grams in x and y sets respectively. The overall similarity of the figures under the text matching dimension is based on the pairwise comparison of the component's texts, which was obtained using the equation (1) above. Therefore, the text matching similarity score is calculated as the sum of the text similarity scores of the matched pair of components, divided by the total number of components in the figures. That is, the text matching similarity between the figures f_1 and f_2 is defined as:

$$Sim_{text}(f_1, f_2) = \frac{2 \cdot \sum_{(n,m) \in M_{Sim}^{opt}} Sim(n, m)}{|N_1| + |N_2|} \dots \dots \dots (2)$$

Where: $Sim(n, m)$ is the similarity score obtain from mapping a pair of components. $|N_1|$ and $|N_2|$ are the number of components in the two figures.

5.2 Structural Similarity Matching

The structural similarity measure is another approach that covers the entire structure of the figure in determining their similarity. It takes into account both the text within the components of the figure as well as the relationship between them. In this approach, the entire figure is considered as a labeled graph. The components and the relationship between them will be considered as nodes and edges of the graph. Then the similarity between the figures can be determined by using graph edit distance [27]. The graph edit distance between the two graphs is the minimum number of the graph edits operations (such as node insertion or deletion, node substitution, edge insertion or deletion and substitutions) necessary to transform one graph to another. This technique is applied considering the following three conditions: The two nodes are considered ‘substituted’ if they mapped. Therefore, their distance is one minus the

similarity of their labels i.e. the text contained in the nodes. The nodes that do not mapped are either deleted or inserted. If there is an edge between two nodes in one graph, then it is expected that such edge should exist in the other graph if and if the nodes are mapped to the nodes in the other graph and there is an edge between the mapped nodes. Otherwise, the nodes are considered deleted. Once the mapping between the nodes and edges are computed for the substituted and inserted or deleted nodes and edges, the graph edit distance can be calculated. The graph edit distance is computed as one the average of the fraction of inserted or deleted nodes, the fraction of inserted or deleted edges and the average distance of substituted nodes. They defined the graph edit distance similarity as:

$$\text{Simged}(q, d) = 1 - \frac{\{snv, sev, sbv\}}{\dots\dots\dots(3)},$$

Where:

$$snv = \frac{|sn|}{|N_1| + |N_2|} \quad sev = \frac{|se|}{|E_1| + |E_2|}$$

$$sbv = \frac{2 \cdot \sum_{(n,m) \in M} 1 - \text{Sim}(n, m)}{|N_1| + |N_2| - |sn|}$$

, $|N_1|$ and $|N_2|$ are the number of nodes in two graphs, and $|E_1|$ and $|E_2|$ are edges contained in two graphs respectively.

5.3 Similarity Calculation Example

In order to see the how each module of the proposed methods work, an example is presented in this section. Consider a source figure and the plagiarized figure in the figures 6(a) and 6 (b) shown below.

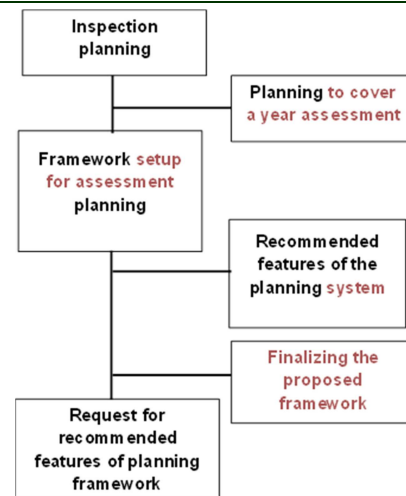


Figure 6 (a): Plagiarized figure

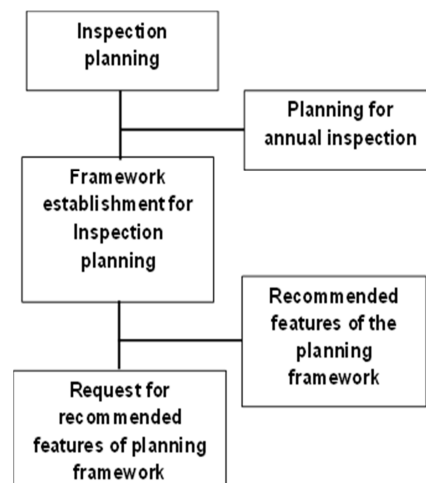


Figure 6 (b): Source figure

Solving by text base similarity computation, the first step is to extract all the text from the two figures and split them as a unique word. For example the word “Operating Expenses” will be extracted as (Operating and Expenses), and the repeated words in each component are counted once. By storing text in a component pair, the two figures can be represented as: $q = \{(\text{inspection, planning}); (\text{planning, cover, year, assessment}); (\text{framework, setup, assessment, planning}); (\text{recommended, features, planning, system}); (\text{request, recommended, features, planning, framework}); (\text{finalizing, proposed, framework})\}$ $d = \{(\text{inspection, planning}); (\text{planning, annual, inspection}); (\text{framework, establishment, inspection, planning}); (\text{recommended, feature,}$

planning, framework); (request, recommended, features, framework, accomplish, inspection).

The second step is similarity scores based on components matching pairs using Jaccard similarity defined as: Assuming the threshold of 0.5, the similarity score of the components included in the matching is: $Sim(n_1, m_1) = 1.0$, $Sim(n_4, m_4) = 0.60$, $Sim(n_5, m_6) = 0.80$. The remaining components are not considered in the matching because there similarity score between all other possible pairs of components is less than the threshold. Finally the overall similarity between the two figures is calculated using the formula:

$$Sim_{text}(f_1, f_2) = \frac{2 * \sum_{(n,m) \in M_{sim}} Sim(n, m)}{|N_1| + |N_2|} = \frac{2 * (1.0 + 0.60 + 0.80)}{5 + 6} = 0.4364$$

That is, the similarity score of the two figures using text matching similarity is equal to 0.4364.

By the structural similarity however, the first step is to calculate similarity between the components. Assuming the same similarity score calculated by the first method, the overall similarity will be calculated using graph edit distance defined as follows:

$$Sim_{ged}(q, d) = 1 - \overline{\{snv, sev, sbv\}}$$

To compute the graph edit distance, the score is computed based on the three graph edit operations which are: $|sn|$: the number of components that are not considered in the matching (that have similarity scores less than the threshold), $|se|$: the number edges that exist between the unmatched components, and $|sb|$: the set of components that matched with other components in the two figures, whose distance is calculated as: $1 - Sim(n, m)$.

Therefore, based on the similarity scores obtain between all the components, $Sim(n_1, m_1) = 1.0$, $Sim(n_4, m_4) = 0.60$, $Sim(n_5, m_6) = 0.80$, it follows that:

$$|sn| = 5, \quad |se| = 9, \quad \text{and} \quad |sb| = 2 * \sum_{(n,m) \in M} (1 - Sim(n, m)) = 2 * (1 - 1 + 1 - 0.6 + 1 - 0.8) = 0.28$$

Based on these parameters, the fraction of each will be calculated as follows:

$$snv = \frac{|sn|}{|N_1| + |N_2|} = \frac{5}{5 + 6} = 0.454 \quad sev = \frac{|se|}{|E_1| + |E_2|} = \frac{9}{4 + 5} = 1$$

$$sbv = \frac{2 * \sum_{(n,m) \in M} (1 - Sim(n, m))}{|N_1| + |N_2| - |sn|} = \frac{2 * 0.83}{5 + 6 - 5} = 0.277.$$

Therefore the overall similarity between the two figures using the structural similarity matching is one minus the average of the three parameters calculated above which is given by:

$$Sim_{ged}(q, d) = 1 - \overline{\{snv, sev, sbv\}} = 1 - 0.55 = 0.45$$

6. Experimental Design

The experiment was conducted with 50 samples of the figure images taken from different sources including the academic projects, examples from text books, scientific papers and internet. The experiment has 35 queries which were manually constructed to cover the seven different instances of plagiarism such as:

- Whole figure plagiarism
- Subcomponents plagiarism,
- Swapping of the components in the source figure,
- Changing connection between the components,
- Text paraphrasing by rewording the contents of the original figure,
- Summarizing the contents of the original figure and
- A combination of Swapping the components and manipulation of the text contents.

Each of plagiarism instances can be queried five times to generate the similarity search, both for the text similarity and structural similarity techniques. Therefore each of the instances have 10 queries, 5 sets for query using text similarity methods and the other 5 sets for query using the structural similarity method. In other word, the experiment has 70 sets of queries for all the seven cases of plagiarism considered in this paper, 35 sets for query using the text similarity methods and 35 sets for query using the

structural similarity methods. However the retrieved results obtained from these sets of queries were used to compute the precisions and recalls which can be used to compare the effectiveness of the two methods for detecting the figure plagiarism.

7. Experimental Results

The results of the experiment are presented based on the two methods defined in this paper. The first method is the text similarity based which is concern about detecting the figure plagiarisms based on matching the texts contained within the figures. The second method is the structural similarity which focuses on recovering the plagiarized figure by taking into account both the text and the relationship similarity between the components of the figures. Based on these methods all the queries were compared with figures in the database and the results were recorded and ranked from the highest to lowest similarity score. The performances of these methods were evaluated using the precision and recall measures derived from their similarity values. Details about the evaluations in terms of similarities and precision and recall for each pattern of the plagiarism are discussed in the following subsections.

7.1 Similarity Result for whole figure Plagiarism

Whole figure plagiarism is a kind of plagiarism where a plagiarist takes the whole image of the figure and makes little alteration on the texts of the source figure. These kinds of alteration on the source figure may not change anything about the semantic meaning of the figures and therefore can be detected easily, such as changing the shape of the figure. To detect this pattern of plagiarism, the query figures that cover this pattern were used to query the figure plagiarism detection system to retrieve the most similar figures in the database. The system retrieves the similar figures with a similarity value between the range of one and zero [0, 1] based on the number of common terms that appear in each sentence (where 0 indicates that the figures are not similar and 1 means exactly the same). The table 4.2 shows the precision and recall for each of the representation methods. From the table 4.2, recall and

precision indicate that the recall and precision of the structural representation are almost the same as that of the text representation for whole figure plagiarism detection. This implies that both representations have the same effects in whole figure search queries when the contents are not so much altered, because the text representation technique ranks a figure higher when the terms in the search figures appear more often.

Table 4.2: Precision and Recall for Whole Figure Plagiarism.

Case	Text Representation		Structural Representation	
	Recall	Precision	Recall	Precision
Whole Figure Plagiarism	0.02	0.50	0.02	0.60
	0.04	0.50	0.04	0.50
	0.04	0.30	0.04	0.50
	0.06	0.30	0.06	0.50
	0.06	0.25	0.06	0.30
Average	0.06	0.42	0.06	0.47

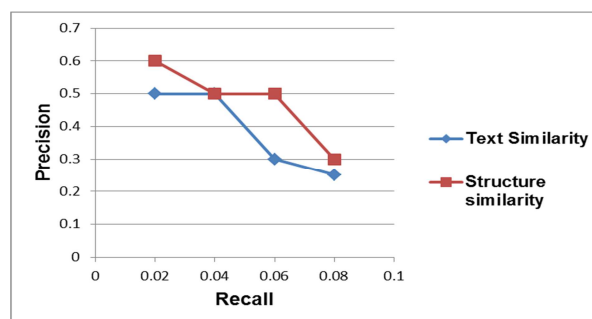


Figure 4.3: Precision and Recall for whole figure plagiarism

7.2 Similarity Result for Subcomponents Plagiarism

Subcomponent plagiarism is a kind of plagiarism where a plagiarist takes a subcomponent of the source figure and makes little or no alteration on the part of the figure being copied. For the queries posted to the figure plagiarism detection system the system generates all the figures that have the same components and highlights the degree of similarity for each method. The table 4.4 shows the precision and recall for each of the representation. The results indicate that both the

text and structural similarities have good performance in this pattern, but the structural similarity has the more effects in subcomponents search query figures when the text contents are slightly altered, because the text representation technique only ranks a figure higher when the terms in the search figures appear more often.

Table 4.2: Precision And Recall For Subcomponents Plagiarism

Case	Text Representation		Structural Representation	
	Recall	Precision	Recall	Precision
Subcomponents Plagiarism	0.02	0.50	0.02	0.60
	0.04	0.48	0.04	0.50
	0.06	0.30	0.06	0.40
	0.06	0.30	0.06	0.30
	0.08	0.25	0.08	0.30
Average	0.052	0.33	0.052	0.42

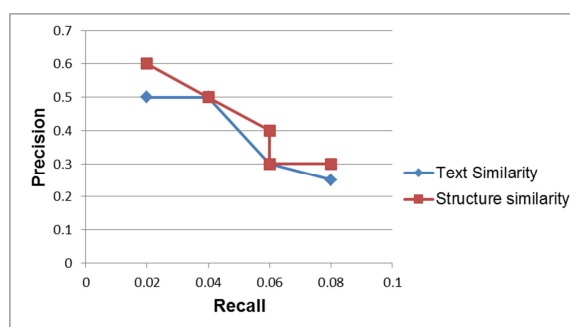


Figure 4.3: Precision And Recall For Subcomponent Plagiarism

7.3 Similarity Result for paraphrasing the Figure Content

Paraphrasing plagiarism is a type of plagiarism where a plagiarist takes the same figure and expresses it using different words in order to obfuscate the plagiarism crime. The effect of this pattern of plagiarism could be detected by comparing the query figures with the set of figures in the data set which were constructed by rewording the source figure in order to detect the proportion of plagiarism by the system. Based on the query, the system retrieved the set of similar figures which were considered similar by the system for both the two methods presented in this paper. The table 4.6 shows the precision and recall for each of the representation. By looking more closely at the results from the table 4.6, the

results indicate that the precision of the structural representation is higher than that of the text representation for the text paraphrasing plagiarism detection. This conformed to the theoretical assumption that components are more likely to match only to the components with strong syntactic similarity when text representation is used. The use of the structural representation has more effect than text representation when the text contents of the figures were paraphrased. The figures 4.9 show the recall and precision graph for text and structural representations for sentence paraphrasing plagiarism detection.

Table 4.6: Precision And Recall For Text Paraphrasing Plagiarism

Case	Text Representation		Structural Representation	
	Recall	Precision	Recall	Precision
Text paraphrasing Plagiarism	0.02	0.60	0.02	0.75
	0.04	0.50	0.04	0.60
	0.04	0.50	0.04	0.60
	0.06	0.30	0.06	0.50
	0.08	0.25	0.08	0.30
Average	0.053	0.39	0.053	0.50

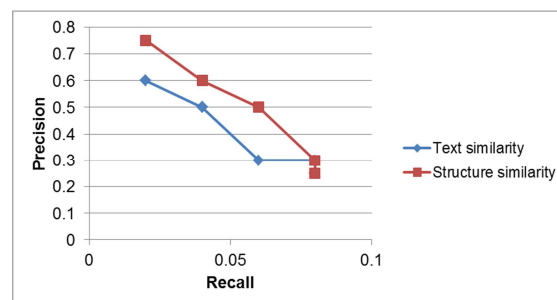


Figure 4.9: Precision And Recall For Paraphrasing Plagiarism

7.4 Similarity Result for Sentence Summarization

Sentence reduction or summarization is another type of plagiarism where a plagiarist takes the idea from figure by summarizing the text in the source figure without referencing the source. To detect this pattern of plagiarism the queries posted the plagiarism detection system

were to suit this type of plagiarism and the system generates all the figures that have common terms based on the degree of similarities for each method. The results from the table 4.8, indicate that the use of the structural representation has more effect than text representation when the text contents of the figures were summarized, based on the same argument when the text were paraphrased. The Figure 4.12 shows the recall and precision graph for text and structural representations for sentence summarization plagiarism detection.

Table 4.8: Precision And Recall For Sentence Summarization Plagiarism

Case	Text Representation		Structural Representation	
	Recall	Precision	Recall	Precision
Sentence summarization Plagiarism	0.02	0.50	0.02	0.75
	0.04	0.50	0.04	0.60
	0.06	0.30	0.06	0.55
	0.06	0.30	0.06	0.50
	0.08	0.25	0.08	0.30
Average	0.052	0.37	0.052	0.54

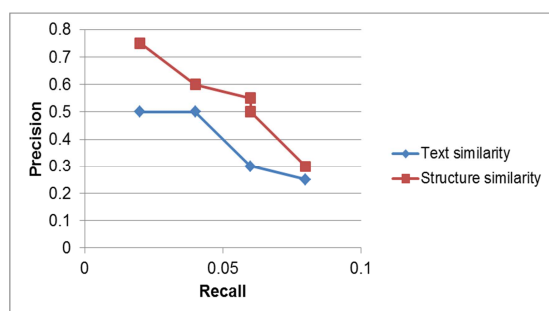


Figure 4.12: Precision And Recall For Text Summarization Plagiarism

7.5 Similarity Result of Changing Connection Plagiarism Detection

Changing connections or the links between the components is another way of plagiarism where a plagiarist takes the whole image of the figure and restructured the shape of the source figure. To detect this pattern of plagiarism, the query figures that cover this pattern were used to query the figure plagiarism detection system to retrieve the most similar figures in the database, and the results of the retrieved figures were

ranked from highest to lowest degree of similarities. The table 4.10 shows the precision and recall for each of the representation. Recall and Precision from the table 4.10, indicate that recall and precision of the text representation are a bit higher than that of the structural representation when the links connecting the components are changed. This is because the text similarity method ranks the figures higher when the terms in the query figures appear more often, and since the structural method considers the links in the similarity estimation even if the text were 100% matched, it ranks the figure less as compared to the text similarity methods. The Figure 4.15 shows the recall and precision for text and structural representations for changing the component plagiarism detection.

Table 4.10: Precision and Recall for Changing Connections between the Components

Case	Text Representation		Structural Representation	
	Recall	Precision	Recall	Precision
Changing Connections Plagiarism	0.02	0.50	0.02	0.60
	0.04	0.50	0.04	0.50
	0.06	0.50	0.06	0.50
	0.06	0.30	0.06	0.25
	0.08	0.25	0.08	0.25
Average	0.06	0.37	0.06	0.39

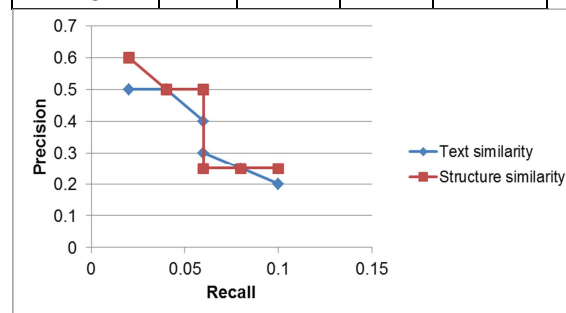


Figure 4.9: Precision And Recall For Changing Plagiarism

7.6 Similarity Result for Swapping of the components

Swapping among the components of the figure is a kind of plagiarism where a plagiarist removes

and substitutes the components by other components from different figures. To detect this pattern of plagiarism, components from different figures were substituted for another in order to test the performance of the system. For the queries posted to the figure plagiarism detection system, it generates all the figures that have the common terms according to their level of similarities. The table 4.12 shows the precision and recall for each of the representation. Results from the table 4.12, indicate that the precision of the structural representation is higher than that of the text representation when the components from different figures were interchange for others. That is to say that the structural similarity has more effect for this pattern of plagiarism because even if the texts within the components are not matched, the links between the unmatched components were counted which add to the similarity scores for the figures being compared. The Figure 4.18 shows the recall and precision graph for text and structural representations for components swapping plagiarism detection.

Table 4.12: Precision And Recall For Swapping The Components Of Figures Plagiarism

Case	Text Representation		Structural Representation	
	Recall	Precision	Recall	Precision
Component Swapping Plagiarism	0.02	0.60	0.02	0.80
	0.06	0.50	0.06	0.50
	0.08	0.25	0.08	0.50
	0.08	0.25	0.08	0.30
	0.10	0.20	0.10	0.25
Average	0.068	0.36	0.068	0.49

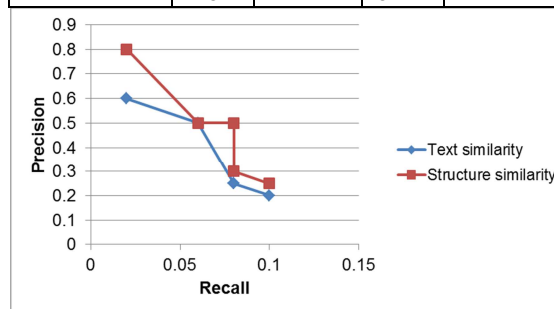


Figure 4.18: Precision And Recall For Swapping Components Plagiarism

7.7 Similarity Result for Swapping Components and Manipulating the Text Contents

This kind of plagiarism is the one in which a plagiarist combines two or more instances of plagiarism by interchanging the components and manipulates the text contents of the figure in order to complicate the plagiarism. These kinds of alteration on the source figure can make detection very complex as this may affect the semantic meaning of the figures. To identify this pattern of plagiarism, the query figures were constructed to suit this pattern and the system retrieved the similar figures and ranked the result according to the degree of similarity. The table 4.14 shows the precision and recall for each of the representation. Although the result of this pattern shows that the text representation retrieves more figures for each query the precision is very low. The structural representation retrieved fewer results but better precision as compared to the text representation. This implies that when the figure is modified by changing the text, the text similarity approach missed out to detect some figures which were actually plagiarized. So the text representation has less effect in detecting this type of plagiarism.

Table 4.14: Precision and Recall for Swapping among the Components and Changing Text Plagiarism

Case	Text Representation		Structural Representation	
	Recall	Precision	Recall	Precision
Swapping among the components and changing text	0.06	0.50	0.06	0.60
	0.08	0.30	0.08	0.50
	0.08	0.25	0.08	0.40
	0.10	0.25	0.10	0.20
	0.12	0.20	0.12	0.20
	0.12	0.20	0.12	0.20
Average	0.09	0.30	0.09	0.38

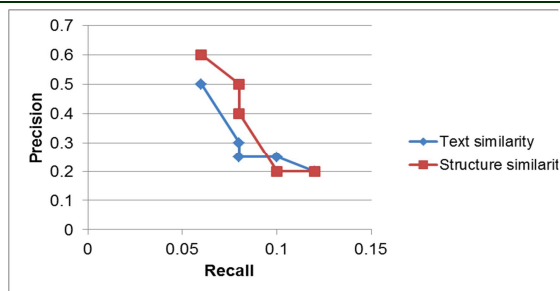


Figure 4.21: Precision And Recall For Swapping And Changing The Text Plagiarism

8.COMPARISONS OF THE TEXT AND STRUCTURAL SIMILARITY

The results achieved by using the text representation are quite different from the results achieved by using the structural representation as shown in the previous sections. Structural representation attains better results because it measures the similarity between the figures based on the text within the components of the figures as well as the relationship between the components, while the text representation measures the similarity only based on the words in the text of the figure components.

The structural representation seems to have a strong effect in detecting plagiarized figures where the text representation is weak. For example when the figure is modified by replacing the words in the components by synonyms or taking a subcomponent of the figure, the structural similarity has more effect than text similarity in detecting such type of plagiarism, because the text similarity technique focuses more on the syntactic similarity of the words. Invariably, this technique degrades more rapidly when the texts were restructured and when the threshold is set high. The advantage of the structural similarity is that even when the contents were restructured as long as there exist a link between the mapped components, the technique takes it into account which would add more effect on their similarity. The table 4.8 shows the average precision and recall across all the pattern of plagiarisms and the figures 4.15 the average precision and recall graph of the table 4.15

Table 4.15: The Average Precision And Recall Across All 7 The Patterns Of Plagiarism

S / N	Patterns of Plagiarism	Text Similarity		Structural Similarity	
		Recall	Precision	Recall	Precision
1	Whole figure plag.	0.053	0.37	0.052	0.48
2	Taking Subcomponent	0.052	0.33	0.052	0.42
3	Text Paraphrasing	0.053	0.39	0.053	0.50
4	Summarizing the text	0.050	0.37	0.052	0.54
5	Changing Connection	0.060	0.38	0.060	0.39
6	Swapping the Comp.	0.068	0.36	0.068	0.47
7	Swapping / manipulate text	0.088	0.30	0.088	0.38
Average		0.061	0.35	0.061	0.45

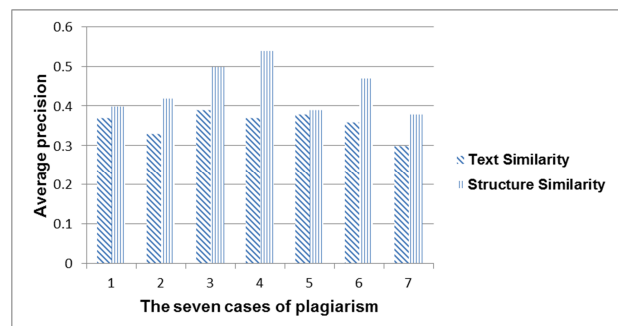


Figure 4.22: The Average Precision And Recall Across The Seven Patterns Of Plagiarism

9 DISCUSSION AND CONCLUSIONS

In this paper we propose two representations that can be used for figure plagiarism detection. From the results of the experiments across all the seven patterns of plagiarism which include: whole figure plagiarism, subcomponent plagiarism, paraphrasing plagiarism, summarization plagiarism, changing connection plagiarism, swapping among the component and swapping the component as well as text manipulation plagiarism; it is observed that structural representation out performed textual

representation. It therefore follows that structural representation is the best of representations which can be able to capture every pattern of plagiarisms. Although the text representation favors detection when plagiarism involves exact copy or verbatim copy of the original figure, such as whole figure plagiarism or taking subcomponent of the figures, it degrades fast when the text contents were extensively modified.

However we have so far focused on developing the techniques for detecting figure plagiarism detection which only considers the syntactic similarity of words, but not the semantic similarity of terms. Therefore there is need to develop more techniques to cover detection of figures whose terms were replaced by their synonyms. Another limitation of this work is that these techniques focused on detecting only standalone figures in academic papers. Work still needed to be done for detecting the figures embedded a scientific document and its source documents.

ACKNOWLEDGMENT

the work is supported by the ministry of higher education (MOHE) and research management centre (RMC) at Universiti Teknologi Malaysia (UTM) under research university grant category (Vot:Q.J130000.2528.07H89).

REFERENCES

- [1] Vinod, K., *Plagiarism-history, detection and prevention*. Hygeia, 2011. **3**.
- [2] Alzahrani, S.M. and N. Salim. *On the use of fuzzy information retrieval for gauging similarity of arabic documents*. in *Applications of Digital Information and Web Technologies, 2009. ICADIWT'09. Second International Conference on the*. 2009. IEEE.
- [3] Lu, X., et al. *Automatic categorization of figures in scientific documents*. in *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*. 2006. ACM.
- [4] Liu, C., et al. *GPLAG: detection of software plagiarism by program dependence graph analysis*. in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006. ACM.
- [5] Arwin, C. and S. Tahaghoghi. *Plagiarism detection across programming languages*. in *Proceedings of the 29th Australasian Computer Science Conference-Volume 48*. 2006. Australian Computer Society, Inc.
- [6] Mann, S. and Z. Frew. *Similarity and originality in code: plagiarism and normal variation in student assignments*. in *Proceedings of the 8th Australasian Conference on Computing Education-Volume 52*. 2006. Australian Computer Society, Inc.
- [7] Son, J.-W., S.-B. Park, and S.-Y. Park, *Program plagiarism detection using parse tree kernels*, in *PRICAI 2006: Trends in Artificial Intelligence 2006*, Springer. p. 1000-1004.
- [8] Ji, J.-H., G. Woo, and H.-G. Cho, *A source code linearization technique for detecting plagiarized programs*. *ACM SIGCSE Bulletin*, 2007. **39**(3): p. 73-77.
- [9] Jiang, L., et al. *Deckard: Scalable and accurate tree-based detection of code clones*. in *Proceedings of the 29th international conference on Software Engineering*. 2007. IEEE Computer Society.
- [10] Zini, M., et al. *Plagiarism detection through multilevel text comparison*. in *Automated Production of Cross Media Content for Multi-Channel Distribution, 2006. AXMEDIS'06. Second International Conference on*. 2006. IEEE.
- [11] Niezgoda, S. and T.P. Way. *SNITCH: a software tool for detecting cut and paste plagiarism*. in *ACM SIGCSE Bulletin*. 2006. ACM.
- [12] Sorokina, D., et al. *Plagiarism detection in arXiv*. in *Data Mining, 2006. ICDM'06. Sixth International Conference on*. 2006. IEEE.
- [13] Stein, B., S.M. zu Eissen, and M. Potthast. *Strategies for retrieving plagiarized documents*. in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 2007. ACM.
- [14] Liu, Y.-T., et al. *Extending Web search for online plagiarism detection*. in *Information Reuse and Integration, 2007. IRI 2007. IEEE International Conference on*. 2007. IEEE.
- [15] Zu Eissen, S.M., B. Stein, and M. Kulig, *Plagiarism detection without reference collections*, in *Advances in data analysis 2007*, Springer. p. 359-366.

- [16] Řehůřek, R., *Text segmentation using context overlap*, in *Progress in Artificial Intelligence* 2007, Springer. p. 647-658.
- [17] Lukashenko, R., V. Gaudina, and J. Grundspenkis. *Computer-based plagiarism detection methods and tools: an overview*. in *Proceedings of the 2007 international conference on Computer systems and technologies*. 2007. ACM.
- [18] Engels, S., V. Lakshmanan, and M. Craig. *Plagiarism detection using feature-based neural networks*. in *ACM SIGCSE Bulletin*. 2007. ACM.
- [19] Godfrey, M.W. and L. Zou, *Using origin analysis to detect merging and splitting of source code entities*. *Software Engineering*, IEEE Transactions on, 2005. **31**(2): p. 166-181.
- [20] Nejati, S., et al. *Matching and merging of statecharts specifications*. in *Proceedings of the 29th international conference on Software Engineering*. 2007. IEEE Computer Society.
- [21] Mehra, A., J. Grundy, and J. Hosking. *A generic approach to supporting diagram differencing and merging for collaborative design*. in *Proceedings of the 20th IEEE/ACM international Conference on Automated software engineering*. 2005. ACM.
- [22] Hassan, M.M. and W. Al Khatib. *Similarity Searching In Statistical Figures Based On Extracted Meta Data*. in *Computer Graphics, Imaging and Visualisation*, 2007. CGIV'07. 2007. IEEE.
- [23] Woo, H.G. and W.N. Robinson. *Reuse of scenario specifications using an automated relational learner: a lightweight approach*. in *Requirements Engineering*, 2002. *Proceedings. IEEE Joint International Conference on*. 2002. IEEE.
- [24] Ehrig, M., A. Koschmider, and A. Oberweis. *Measuring similarity between semantic business process models*. in *Proceedings of the fourth Asia-Pacific conference on Conceptual modelling-Volume 67*. 2007. Australian Computer Society, Inc.
- [25] Minor, M., A. Tartakovski, and R. Bergmann, *Representation and structure-based similarity assessment for agile workflows*, in *Case-Based Reasoning Research and Development* 2007, Springer. p. 224-238.
- [26] B. SRISURA, *RETRIEVING USE CASE DIAGRAM WITH CASE-BASED REASONING APPROACH*. *Journal of Theoretical and Applied Information Technology*, 2005.
- [27] Dijkman, R., et al., *Similarity of business process models: Metrics and evaluation*. *Information Systems*, 2011. **36**(2): p. 498-516.
- [28] Blok, M.C. and J.L. Cybulski. *Reusing uml specifications in a constrained application domain*. in *Software Engineering Conference, 1998. Proceedings. 1998 Asia Pacific*. 1998. IEEE.
- [29] Udomchaiporn, A., N. Prompoon, and P. Kanongchaiyos. *Software Requirements Retrieval Using Use Case Terms and Structure Similarity Computation*. in *Software Engineering Conference, 2006. APSEC 2006. 13th Asia Pacific*. 2006. IEEE.
- [30] Minkov, E. and W.W. Cohen. *Learning graph walk based similarity measures for parsed text*. in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2008. Association for Computational Linguistics.
- [31] Wombacher, A., *Evaluation of technical measures for workflow similarity based on a pilot study*, in *On the Move to Meaningful Internet Systems 2006: CoopIS, DOA, GADA, and ODBASE* 2006, Springer. p. 255-272.
- [32] Madhusudan, T., J.L. Zhao, and B. Marshall, *A case-based reasoning framework for workflow model management*. *Data & Knowledge Engineering*, 2004. **50**(1): p. 87-115.
- [33] Melnik, S., H. Garcia-Molina, and E. Rahm. *Similarity flooding: A versatile graph matching algorithm and its application to schema matching*. in *Data Engineering, 2002. Proceedings. 18th International Conference on*. 2002. IEEE.
- [34] Madhavan, J., P.A. Bernstein, and E. Rahm. *Generic schema matching with cupid*. in *Proceedings of the International Conference on Very Large Data Bases*. 2001.
- [35] Osman, A.H., N. Salim, and M.S. Binwahlan, *Plagiarism Detection Using Graph-Based Representation*. arXiv preprint arXiv:1004.4449, 2010.
- [36] Fathi Taibi, F.M.A., Md. Jahangir Alam *Similarity Detection in Collaborative Development of Object-Oriented Formal Specifications*. *World Academy of Science, Engineering and Technology* 21 2008, 2008.