IMPROVED SEMANTIC GRAPH-BASED PLAGIARISM DETECTION

AHMED HAMZA OSMAN AHMED

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

MARCH 2013

To my beloved parents, brothers, sisters, wife, son (Albraa), and daughter (Braa'h)

ACKNOWLEDGEMENT

First of all, I would like “Praise be to Allah, the cherisher and the sustainers of the world”, “praise be to him he who taught by the pen, taught man, that which he did not know”

First of all, I would like to thank my supervisor professor Dr. Naomie Binti Salim for allowing me to carry out this thesis, and the wonderful experience and all the good times she gave to me in addition to, continuous motivation and encouragement. I also owe my great deal of thank to UTM and Faculty of Computer Science and Information System for providing the good environment to carry out the research, and to our professors and staff for their friendly support and help throughout the study and special thank to Prof. Dr. Naomie for supporting me in this study.

And most importantly, I would like to thank to my family for their love and support, especially my mother, my father and my wife for their continuous support and supplication.

This thesis would not see the light and would not have been possible without the continuous financial support of my university “International University of Africa” and for providing me the opportunity to study in Malaysia.

Finally, I would like to thank all my friends and colleagues for their support and help.

ABSTRACT

Plagiarism detection occurs when the content of a text is copied without permission or citation. Nowadays, many text documents on the internet are easily copied and accessed. This study proposed improved methods to handle plagiarism. The proposed plagiarism detection methods are developed using graph-based representation and semantic role labeling which are improved using fuzzy logic technique and chi-squared automatic interaction detection. The graph-based method does not only represent the content of a text document as a graph, but also captures the underlying semantic meaning in terms of the relationships among its concepts. Semantic role labeling is superior in generating semantic arguments for each sentence. This semantic role labeling plays an important part in plagiarism detection as it segments the role of concepts in documents to labels which are compared and used to detect plagiarism. Scoring for each argument generated by the fuzzy logic method to select important arguments is also another feature of this study. Chi-squared Automatic Interaction Detection technique was applied to enforce the results obtained from the fuzzy logic and semantic role labeling by selecting important arguments from the sentences. It is concluded that not all arguments in the text are useful in the plagiarism detection process. Therefore, only the most important arguments were selected by the fuzzy logic and Chi-squared automatic interaction detection, and the results were used in the similarity calculation process. Experiments were tested on the PAN-PC-2009 for standard artificial simulation corpus and the Short Answers Questions (CS11) for human simulation corpus in plagiarism detection. The proposed methods detected many types of plagiarisms, such as copy paste plagiarism, rewording or synonym replacement, changing of word structure in the sentences, modifying the sentence from passive voice to active voice and vice-versa. Results from the experiments using the proposed methods in comparison to other palagiarism detection techniques (Fuzzy Semantic-Based String Similarity and Longest Common Subsequence) achieved better performance in terms of recall (93%), precision (90%) and *f*-measure (91%).

ABSTRAK

Pengesanan plagiat berlaku apabila kandungan dalam teks disalin tanpa kebenaran atau tidak diberi rujukan. Kini, kebanyakan dokumen teks di internet mudah disalin dan dicapai. Kajian ini mencadangkan kaedah diperbaiki untuk menangani masalah plagiat. Kaedah pengesanan plagiat yang dicadangkan dibangunkan dengan menggunakan perwakilan berasaskan graf dan pelabelan peranan semantik yang kemudiannya diperbaiki dengan teknik logik kabur dan *Chi-squared Automatic Interaction Detection*. Kaedah berasaskan graf tidak mewakilkan kandungan dokumen teks secara graf sahaja, tetapi juga boleh mendapatkan makna semantik dari segi hubungan antara konsep-konsepnya. Pelabelan peranan semantik adalah sangat berkesan untuk menjana peranan-peranan semantik bagi setiap ayat. Ia memainkan peranan penting untuk mengesan plagiat dengan membahagikan peranan-peranan semantik bagi konsep-konsep dalam dokumen di mana plagiat dapat dikesan melalui perbandingan label. Penskoran setiap peranan semantik menggunakan kaedah logik kabur juga merupakan satu lagi ciri kajian ini bagi tujuan memilih peranan-peranan semantik yang penting. Teknik *Chi-squared Automatic Interaction Detection* juga digunakan untuk memperbaiki keputusan yang diperolehi daripada logik kabur dan pelabelan peranan semantik untuk memilih peranan-peranan semantik penting daripada ayat. Adalah didapati bahawa tidak semua peranan semantik dalam teks penting kepada proses pengesanan plagiat. Oleh itu, hanya peranan-peranan semantik yang penting sahaja dipilih oleh logik kabur dan *Chi-squared Automatic Interaction Detection*, di mana keputusan yang diperolehi melaluinya digunakan bagi proses pengiraan persamaan. Eksperimen bagi kajian ini dijalankan menggunakan data dari PAN-PC-2009 untuk korpus simulasi tiruan dan juga data dari *Short Answers Questions (CS11)* untuk korpus simulasi manusia bagi pengesanan plagiat. Kaedah yang dicadangkan boleh mengesan pelbagai jenis plagiat, seperti salin dan tampal, penukaran perkataan atau penggantian sinonim, perubahan struktur perkataan dalam ayat, pengubahsuaian ayat dari ayat pasif kepada ayat aktif dan sebaliknya. Keputusan yang dicapai daripada ujikaji yang dijalankan menunjukkan teknik-teknik yang dicadangkan memberi prestasi yang lebih baik dari segi dapatan semula (93%), keperisian (90%) and ukuran-*F*(91%) berbanding dengan teknik-teknik lain dalam pengesanan plagiat seperti *Fuzzy Semantic-Based String Similarity* dan *Longest Common Subsequence.*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

α	-	Function Labelling The Nodes
B	-	Function Labelling The Edges
A_F	-	Arguments features
ADV	-	General purpose
ALG	-	Argument Label Group
Arg0	-	Agent
Arg1	-	Direct Object/Theme/Patient
Arg2–5	-	Arguments not fixed
Bellmf	-	Bell Membership Function
BOW	-	Bag of Word
CART	-	Classification and Regression Trees
CHAID	-	Chi-squared Automatic Interaction Detection
CHK	-	Chunking
CoNLL	-	Conference on Natural Language Learning
COPS	-	Copy Protection System
CS11	-	Clough and Stevenson Corpus
DIR	-	Direction
DIS	-	Discourse Connectives
E	-	Set of edges connecting with nodes
E	-	Edge
EXT	-	Extent
FLIS	-	Fuzzy Logic Inference System
FLS	-	Fuzzy Inference System
FM	-	Frequency Model
G	-	Graph, Data Graph
Gaussimf	-	Gaussian Membership Function

HITS	-	Hyperlink Induced Topic Search
ID	-	Identification Number
IR	-	Information Retrieval
LCS	-	Longest common subsequence
LOC	-	Location
LSI	-	Latent Semantic Indexing
LZ	-	Lempel-Ziv
MCS	-	Minimum common supergraph
Mcs	-	Maximum common subgraph
ML-SOM	-	Multi-Layer Self-Organizing Maps model
MMCSN	-	Maximum Minimum Common Subgraph/ Supergraph Normalized Measure
MNR	-	Manner
MOD	-	Modal verb
NEG	-	Negation marker
NER	-	Name Entity Recognition
NLP	-	Natural Language Processing
O	-	Adjective
P	-	Pattern Graph
PAN-PC	-	PAN Plagiarism Corpus
PNC	-	Purpose
POS	-	Part-of-Speech
Propbank	-	Proposition Bank
P-value	-	Probability Value
RFM	-	Relative Frequency Model
SC	-	Structural Characteristic
SCAM	-	Stanford Copy Analysis Method
SIM	-	Total similarity score values between the original and suspected documents
SPT	-	Stanford Parser Tree
SRL	-	Semantic Role Labelling
STA	-	Semantic Term Annotation
TF-IDF	-	Term Frequency–Inverse Document Frequency

TMP	-	Time
Trap	-	Trapezoidal membership function
Trimf	-	Triangular Membership Function
TSK	-	Takagi-Sugeno-Kang
TTS	-	Chinese to Taiwanese System
UK	-	United Kingdom
USA	-	United States of America
V	-	Set of nodes (vertices)
V	-	Node or Vertex
V	-	Verb
VGj	-	Vertex Mapping of Data Graph
Vpi	-	Vertex Mapping of Pattern Graph
VSM	-	Vector Space Model

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Plagiarism is defined as the "unacknowledged copying of documents or programs" (Joy and Luck, 1999). It can occur in many areas. For example, companies may look for a competitive advantage in the market, and academicians may need to advance their careers by way of quick publishing. Many empirical studies and analyses have been undertaken by the academic community to deal with student plagiarism. The correct selection of text features is a key aspect in the task of discriminating between plagiarized documents and non-plagiarized documents.

There are many types of plagiarism mentioned by Hermann, Frank, and Bilal,(2006), such as copy and paste, redrafting or paraphrasing of the text, plagiarism of ideas, and plagiarism through translation from one language to another.

Nowadays, many documents are available on the internet and are easy to access. Due to this wide availability, users can easily create a new document by copying and pasting. Sometimes users can reword the plagiarized part by replacing words with their synonyms. This kind of plagiarism is difficult to be detected by using traditional plagiarism detection systems such as copy protection system

(COPS) (Sergey *et al.*, 1995a), Stanford Copy Analysis Method (SCAM) (Sergey *et al.*, 1995b) or CHECK (Antonio *et al.*, 1997).

The biggest challenge in plagiarism detection is to provide plagiarism checking with appropriate algorithms in order to improve the odds of finding instances of plagiarism and to decrease the time spent checking. According to the literature reviewed by Alzahrani, Salim and Abraham (2011), and Bao (2003), current plagiarism detection systems were found to be too slow due to the matching techniques such as string and character matching, and the matching algorithms were dependent on the text's lexical structure rather than its semantic structure, making it difficult to detect any text that had been paraphrased. The important question for the plagiarism detection problems examined in this study is whether application of new techniques such as Graph-based, Semantic Role Labelling (SRL), Fuzzy Logic and Chi-squared Automatic Interaction Detection (CHAID) algorithms can improve plagiarism detection of texts.

In this study, we propose a new plagiarism detection methods based on Graph-based and Semantic Role Labelling (SRL). We later improved these methods using Fuzzy Logic and CHAID algorithm. The proposed method can detect copy and paste plagiarism, rewording or synonym replacement, changing of word structure in the sentences, as well as modification of the sentence from passive voice to active voice and vice versa.

The organization of this chapter is described as follows: Section 1.2 reviews the background of the problem while section 1.3 will present the problem statement. Then, Section 1.4 will discuss the objective of this study. Section 1.5, 1.6, and 1.7 will focus on the scope of study, significance of study and expected contribution, respectively. Finally, Section 1.8 will describe the organization of this thesis.

1.2 Background of Research

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. Plagiarism means using the text written by others which may be re-adjusted by adding or deleting text but without any citation or reference to the original author.

There are many types of plagiarism, such as copy and paste, redrafting or paraphrasing of the text, plagiarism of ideas, and plagiarism through translation from one language to another. According to Adeva, Carroll and Calvo, (2006), at least 10% of student's work is likely to be plagiarized in USA, Australia and UK universities (Lyon *et al.*, 2006). Another current research project found that 70% of students confess to small instances of plagiarism, and about half of the students studied were guilty of a cheating offence on a written assignment. Additionally, 40% of students confess to using the cut and paste method when completing their assignments (McCabe, 2005).

Differentiating between the plagiarized documents and non-plagiarized documents in an effective and efficient way is one main issue in the field of plagiarism detection. Current methods of plagiarism detection are based on character matching, n-gram, chunks or terms (Alzahrani, *et al.*, 2011; Hermann, *et al.*, 2006; Potthast *et al.*, 2010).

Research institutions and universities require new technology for detecting plagiarism. This is crucial in controlling and marking researchers and students' essays, homework, papers and reports. Many plagiarism detection tools use character matching or string matching method to detect the plagiarized text (Hermann, *et al.*, 2006; Maurer *et al.*, 2006). However, most of the current software and techniques are not effective in detecting plagiarized text because these tools tend to compare a suspected text with original text using characters matching, some by chunks while others by words. This leads to an exhaustive, time consuming search

(Mozgovoy, 2007). As noted above, the purpose of this study is to propose new techniques for plagiarism detection based on Graph-based Representation, Semantic Role Labelling, Fuzzy Logic technique and Chi-squared Automatic Interaction Detection method.

Text Graph-Based Representation does not only represent the content of a text document as a graph, but it also captures the underlying semantic meaning in terms of the relationships among its concepts. Semantic Role Labelling (SRL) is superior in generating arguments for each sentence semantically (Martha Palmer *et al.*, 2005). WordNet Thesaurus (Fellbaum, 1998) is an excellent tool for extracting the concepts or synonyms for each word in the sentences. Fuzzy Logic is a common expert system application which has proven successful in several predictions and control systems (Wong and Hamouda, 2000). Chi-squared Automatic Interaction Detection algorithm (CHAID) is highly visual and simple to understand and interpret. It uses multi-way splits by default; it needs rather large sample sizes to work effectively, since with small sample sizes the respondent groups can become smaller for reliable analysis (Diepen and Franses, 2006). We employed CHAID to detect an interactions between variables in the plagiarism detection corpus. Using this method it is possible to establish relationships between a 'dependent variable' for example, the relationship between total similarity score and other arguments types variables, such as: subject, object, verb, among others. CHAID does this by identifying discrete groups of respondents and, by taking their responses to explanatory variables, seeks to predict what the importance and impact arguments variables will be on the total similarity score variable.

Graph Based Representation relies on the different levels of processing in the text. Zhang (2009) divided graph representation into three levels; the document level, the sentence level and the term level. The representation of these levels within a graph defines the graph node and graph edge. The node can hold either a document or a sentence or a term and the edge is the weight between these levels. The 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. Prominent examples for document

graph-based representation include a network citation and World Wide Web network. In the network citation the authors can cite from the others' work by referring to their work. Each author's paper contains many references while every reference represents a paper or an article or a document. The relationship among these references and author's paper is that the topics covered in the author's paper are similar to those in the references. Due to that, each reference is represented as a node and is linked with the main paper that includes it as references by the edge.

Semantic structures or case frames were introduced by Minsky (1974) where common frames were used for common roles and themes such as FrameNet proposed by Baker , Fillmore and Lowe (1998) and PropBank proposed by (Martha Palmer, *et al.*, 2005). A statistical system is trained on the data from the FrameNet project to automatically assign semantic roles (Daniel Gildea and Daniel Jurafsky, 2002). Surdeanu *et al.*, (2003); Pradhan *et al.*, (2004); and Xue, Nianwen and Martha Palmer, (2004) (Xue, *et al.*, 2004) followed this approach by improving sets of features and machine learning methods. Semantic Role Labelling (SRL) by Johansson and Nugues (2008) achieved the best result in terms of *F*-measure for the corpus evaluation. Barnickel *et al.*,(2009) introduced a large scale application of neural network based on Semantic Role Labelling for automated relation extraction from biomedical texts. This method mainly used SENNA software (Collobert and Weston, 2007). SENNA can extract the arguments of the sentences and semantic role for the terms based on neural network algorithms where the users can adopt this software to extract the semantic relations between terms in text documents.

Semantic Role Labelling is a process used to identify and label arguments in a text. The underlying idea is that the sentence level semantic analysis of text determines the object and the subject of a text. It can be extended to the characterization of events such as determination of "*who"* did *"what"* to *"whom," "where," "when,"* and *"how.*" The predicate of a clause (usually a verb) establishes "what" took place, and other parts of the sentence express the other arguments of the sentence (such as "who" and "when"). The primary task of Semantic Role Labelling is to identify what semantic relation holds among a predicate and its associate participants or properties, with these relations drawn from a pre-defined list of

possible semantic roles for that predicate or class of predicates. The typical labels used in SRL are an Agent, Patient and Location for the entities participating in an event. Those labels can be extended to more specific arguments such as Time and Place in some text.

Currently, there are several tools using Semantic Role Labelling including the Proposition Bank or Propbank (Martha Palmer, *et al.*, 2005), FrameNet (Baker, *et al.*, 1998) and VerbNet (Karin Kipper *et al.*, 2000). PropBank has thus far obtained the attention of the researchers in SRL technique.

Most matching algorithms are dependent on the text's lexical structure rather than its semantic structure (Maurer, *et al.*, 2006). Therefore, it is difficult to detect texts that are paraphrased semantically. Based on Maurer, Kappe, and Zaka (2006), the relation between sentences and their semantic content in plagiarism detection based on text semantic analysis method can be solved by the Semantic Role Labelling (SRL). A semantic role is the underlining relationship that a participant has with the main verb in the clause (Payne, 1997); also known as thematic role, semantic case, the theta role (generate grammar), and deep case (case grammar). SRL is a new method in plagiarism detection and it is superior for generating arguments for each sentence semantically.

An improvement to current plagiarism detection methods proposed by this study can be illustrated by using the Graph Based Representation and Semantic Role Labelling for text documents and then selecting the important arguments that can have an effect on plagiarism detection using Fuzzy Logic method and the CHAID algorithm.

Fuzzy Logic (Zadeh, 1965) is a common expert system applications which has proven successful in several predictions and control systems (Wong and Hamouda, 2000). It is usually used to represent ambiguous and vague information. It is an appropriate method for determining the relationship between inputs and the desired outputs of a system. It has the ability to control decision-making based

on input values; that is, it works under various assumptions and approximations. Fuzzy Logic Inference Systems (FLIS) include some inputs and outputs with a set of predefined rules and a defuzzification process.

Fuzzy Logic was used to control a simple laboratory steam engine (Mamdani, 1974). It is a mathematical assumption of ambiguous reasoning that allows it to obtain results and decision-making model in linguistic terms. Fuzzy Logic has become one of the main successful technologies in many applications and sophisticated control systems. For example, Munakata and Jani (1994) mentioned that over a thousand industrial commercial fuzzy applications have been successfully developed in recent years.

Fuzzy Logic resulted in the development of the theory of fuzzy sets. Classical Logic is limited and can only deal with two values – true or false. However, there is a need for a system that can handle partial truths (neither completely true nor completely false). Fuzzy logic, therefore, is an extension classical logic as it generalizes classical logic inference rules which have the ability to deal with approximate reasoning (Klir and Yuan, 1995).

Some research has been done with fuzzy-set with plagiarism detection, for example the study carried by Alzahrani and Salim (2010); Yerra (2005). Matching fragments of text, such as terms and sentences, become ambiguous or approximate, and applies a range of similarity values from one (fully matched) to zero (totally different). In plagiarism detection filed, the concept of fuzzy can be represented by considering each term in a text is associated with a fuzzy set that comprises terms with the same meaning, and there is a degree of similarity between terms in a text and the fuzzy set (Yerra and Ng, 2005). Fuzzy-set Information Retrieval (IR) for plagiarism detection is effective since it detects not only exact matches but also similar statements based on the degree of similarity between words in the statement and their fuzzy sets. In order to construct the fuzzy-set and the degree of similarity between words, term-to-term correlation matrix should be constructed before using the fuzzy set Information Retrieval (IR). It should contain the words and their

corresponding correlation factor that measure the degree of similarity (degree of membership between 0 and 1) among different word, such as "vehicle" and "transport." The fuzzy-set IR technique obtains the degrees of similarity among sentences by computing the correlation factors between any pair of words from two different sentences in their respective documents. Therefore, fuzzy set IR is capable of not only detecting general similarities but also similar patterns between two documents.

The fuzzy set is an elaboration of the traditional set "crisp set" in which each member has a degree of membership to that set as determined by a membership function. The membership function is a function that assigns a membership degree to each member in the target set and the range of membership degree is between zero and one. The computer can translate the linguistic statement into actions based on a set of "IF-THEN" rules of the Fuzzy Logic. The fuzzy IF-THEN rules are normally created in the form of "if *A* then *B*" in which the condition is connected with actions where *A* and *B* are fuzzy sets. Fuzzy Logic has an advantage in terms of simplicity of development and modification because the rules are well understandable and easy to modify, add new rules, or remove existing rules. One of the objectives of this study is to select the best and most important arguments that can define the plagiarism process and improve the detecting similarity score. Arguments defined as unimportant using FIS will be ignored.

Chi-squared Automatic Interaction Detection or CHAID is an extremely effective predictive statistical method, developed by Kass (1980) used for segmentation. CHAID works by using combining predictor features according to a value that was derived by using statistical test criteria. CHAID then combines values that are similar to the target variables with the remaining, dissimilar values.

CHAID creates a decision tree by using the best predictor to form the first level. Each child node is created, or grown, from a group of features with similar features, like leaves springing from the branches of a tree. This process continues

until the tree has fully grown. Significant statistical tests used stops upon the measurement level of the target field.

Unlike methods that rely on binary trees, the CHAID tree grows wider because it has the ability to use any type of variable and it can use both weight variables and frequency variables.

One of the advantages of CHAID algorithms is that its results are highly visual and simple to understand and interpret (Merel van Diepen and Philip Hans Franses, 2006). Yih-Jeng, Ming-Shing and Chin-Yu, (2008) reports in their research that:

> " A CHAID method has many advantages of applying in looking for patterns in complicated datasets. The level of measurement for the dependent variable and predictor variables can be nominal, ordinal, or interval. The predictor variables need not all be measured at the same level (nominal, ordinal, and interval). For the case of missing values in predictor variables, it can be treated as a "floating category" so that partial data can be used whenever possible within the tree.
>
> (Yih-Jeng, Ming-Shing and Chin-Yu, 2008: 366)

Due to the advantages of The CHAID algorithm, we found that it is very good at improving the plagiarism detection results for our proposed methods. The main reason that the CHAID algorithm was employed in our study was that the algorithm can select important features from the data set and adopted this attribute to select important arguments from the sentences.

CHAID is both a statistical model and data mining method. It is belongs to a set of models identified by decision trees. It is used for classification and prediction purposes. In plagiarism detection, we will employ the CHAID algorithm to predicate important and unimportant arguments from the suspected and original documents.

The arguments will be extracted first from the documents based on SRL. Then, the similarity will be calculated based on the Jaccard similarity measure (Jaccard and Paul, 1901) between the arguments. The input of CHAID algorithms will be a set of arguments similarity scores between original and suspected documents and the output of CHAID will be a set of predicated important arguments. One of the advantages of CHAID algorithm is to select the important features form a set of features. In our plagiarism detection method, we adopted the CHAID algorithm to select the important arguments from a set of the extracted arguments using SRL method. The CHAID algorithm usually predicates the important features in form of decision tree, in the same case; the important arguments will be selected in the tree form.

In addition to Fuzzy Logic and The CHAID algorithm, another improvement that can be made to current plagiarism detection techniques can be found in the mechanism that compares the corresponding arguments between two texts, (Subject with Subject, Verb with Verb) which differs from the traditional comparison mechanisms (Subject with Verb, Subject with Adverb, Subject with Adjective and so on). The greatest benefit of employing this type of plagiarism detection is that it can detect copy and paste, semantic plagiarism, rewording or synonym replacement, changing of word structure in the sentences, modifying the sentence from passive voice to active voice and vice versa.

This study implements methods that detects plagiarism by representing text as graph and selecting arguments based on similarity score using important arguments and adjusting the weighting for each argument to improve the total results using Fuzzy Logic and CHAID algorithm with Semantic Role Labelling to extract key arguments of the original and suspected texts and estimating the relevance of suspected sentences by capturing the main content and the semantic content available in sentences using Graph-based and Semantic Role Labelling.

1.3 Problem Statement

One of the important algorithms in plagiarism detection is the matching algorithms. Matching algorithms focus on the text lexical structure rather than semantic structure. As a result, it is difficult to detect any text paraphrased semantically (Alzahrani, *et al.*, 2011; Hermann, *et al.*, 2006; Potthast, *et al.*, 2010). Additionally, current plagiarism detection systems were also found to be too slow and takes too much time for each checking (Mozgovoy, 2007).

This research is concerned about a semantic matching algorithm to answer the following research question:

(i) *Can graph-based representation of text documents be used for plagiarism detection?*

(ii) *Can the semantic role labelling (SRL) with the graph-based representation enhance plagiarism detection?*

(iii) *Can the arguments weight adjustment to select the important arguments in sentences be used to improve graph-based with SRL plagiarism detection?*

(iv) *How can fuzzy logic and Chi-squared Automatic Interaction Detection algorithm (CHAID) define important arguments that need to be used for plagiarism detection?*

(v) *Can the combination of fuzzy logic and Chi-squared Automatic Interaction Detection algorithm (CHAID) algorithm with semantic role labelling and graph-based give better plagiarism detection technique?*

1.4 Objectives

The aim of this research is to show how to combine a graph-based method, SRL, Fuzzy Logic, and The CHAID algorithm to form new techniques to detect plagiarism. To achieve this goal, the following objectives will be aimed:

1- To develop a graph-based technique that can be used to represent text documents as graph for plagiarism detection.

2- To investigate the use of Semantic Role Labelling (SRL) in the graph-based representation to enhance plagiarism detection.

3- To identify important arguments that can be used to improve detection of plagiarism using weight arguments, fuzzy logic technique, and the CHAID algorithm.

1.5 Research Scope

The preceding section mentioned the objectives of this research which focus on how to produce a good plagiarism detection algorithm. The following aspects are the scope of study for the mentioned objectives.

1- Plagiarism detection using graph-based representation.

2- An improved plagiarism detection based on semantic role labelling.

3- An improved plagiarism detection based on fuzzy semantic role labelling and CHAID algorithm.

4- Evaluation of the performance of the proposed methods using PAN-PC-09 and short answers questions plagiarism corpus and compare with other approaches such as fuzzy semantic-based string similarity, longest common subsequence (LCS), and semantic-based similarity.

1.6 Significance of Research

Much of the research done in the plagiarism detection field is based on character matching method and fingerprint method. This study will introduce plagiarism detection methods which use Graph-based Representation and SRL with Fuzzy Logic and CHAID technique. The text graph-based representation used to represent the content of a text document as a graph and used also to capture the underlying semantic meaning in terms of the relationships among its concepts. SRL was used to analyze the sentences semantically and the WordNet Thesaurus was used to extract the concepts or synonymies for each word inside the sentences. Fuzzy Logic technique focused on the fuzziness of argument terms in the sentence and used it as an optimisation technique with a CHAID algorithm to select important sentence arguments.

1.7 Study Contributions

The expected contribution of this study can be explained as follows:

1. A new plagiarism detection method using graph to represent text document with semantic meaning in terms of the relationships among its concepts. WordNet Thesaurus will be used to extract concepts or synonymies for each word inside the sentences.

2. A new plagiarism detection scheme based on Semantic Role Labelling which can compare sentences semantically.

3. An improved plagiarism detection method based on Arguments Weighting Scheme.

4. A new plagiarism detection method using Fuzzy Logic and CHAID algorithm

1.8 Thesis Organization

This thesis organized with eight chapters. These chapters are as follows:

Chapter1: ***Introduction*****:**

The introductory chapter of this thesis provides a brief overview of some of the issues that of concern to those working in the field of plagiarism detection. This chapter will also look at the goals, the scope of this study as well as examining the contributions this research can make to the field of plagiarism detection.

Chapter 2: ***Literature Review:***

This chapter evaluates state-of-the-art approaches in plagiarism detection. Techniques such as Semantic Role Labelling, Fuzzy Logic, and the CHAID algorithm will be reviewed. In addition, a plagiarism detection evaluation measurements and datasets will be covered too.

Chapter 3: ***Methodology:***

This chapter describes the methodology and principal experiments used to obtain the objectives of this research study. Topics will include; Text Graph-based Representation, Semantic Role Labelling, Fuzzy Logic and Chi-squared Automatic Interaction Detection.

Chapter 4: ***Plagiarism Detection Using Graph-Based Representation Detection:***

This chapter proposes a plagiarism detection technique based on graph based representation. In this method, a text document is represented as a graph and used to captures the underlying semantic meaning in terms of the relationships among its concepts. The comparison between the documents is calculated according to the similarity between the terms and concepts of the sentences.

Chapter 5: ***Plagiarism Detection Based on Semantic Role Labelling:***

This chapter introduces a plagiarism detection method using SRL. SRL is used to analyze the sentences semantically and WordNet Thesaurus is used to extract the concepts or synonymies for each word inside the sentences. This method can detect a plagiarism after terms arguments are extracted. The comparison is calculated according to the semantic position of the terms in the sentences. In addition, this chapter will also propose an improved plagiarism detection scheme based on semantic role labelling conducted by using and argument weight scheme. Arguments behaviours will be studied to select important arguments that can have an effect on plagiarism detection. This improvement reflects the important arguments that should be used in comparison process rather than all extracted arguments using SRL.

Chapter 6: ***An Improved Plagiarism Detection Technique Based On Fuzzy Semantic Role Labelling:***

This chapter introduces an improvement of SRL plagiarism detection technique using Fuzzy Logic. Fuzzy Logic will be used as an optimisation technique by selecting the important arguments in a sentence. Selection for each argument generated by the Fuzzy Logic in order to select important arguments will also be discussed.

Chapter 7: ***An Improved Plagiarism Detection Method Based On Semantic Role Labelling and Chi-Squared Automatic Interaction Detection:***

This chapter introduces an improvement of text similarity checking and plagiarism detection method based on a Semantic Role Labelling (SRL) and Chi-squared Automatic Interaction Detection (CHAID).

Chapter 8: ***Conclusion and Future Work:***

Chapter 8 will review the conclusions of the research discussed throughout this study. This section will also put forward recommendations for future studies.

REFERENCES

Adeva, J.J.G. Carroll, N.L., and Calvo, R.A. (2006). *Applying plagiarism detection to engineering education*.

Ahmad, H. (2006). *Plagiarism Detection Systems: An Evaluation Of Several Systems*.

Akiva, N. (2011). *Using Clustering to Identify Outlier Chunks of Text.* Paper presented at the Lab Report for PAN at CLEF(2011).

Aleman-Meza, B. Halaschek-Wiener, C. Sahoo, S. Sheth, A., and Arpinar, I. (2005). *Template based semantic similarity for security applications*. Intelligence and Security Informatics, 1049-1061.

Alzahrani, S. Palade, V. Salim, N., and Abraham, A. (2011a). *Using structural information and citation evidence to detect significant plagiarism cases in scientific publications*. Journal of the American Society for Information Science and Technology, Vol 63(2), 286-312.

Alzahrani, S., and Salim, N. (2010). *Fuzzy Semantic-Based String Similarity for Extrinsic Plagiarism Detection*. Lab Report for PAN at CLEF(2010).

Alzahrani, S.M. Salim, N., and Abraham, A. (2011b). *Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods*. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, PP(99), 1-1.

Aniruddha Ghosh Pinaki Bhaskar Santanu Pal, and Sivaji Bandyopadhyay. (2011). *Rule Based Plagiarism Detection using Information Retrieval.* Paper presented at the Notebook for PAN at CLEF 2011. In Notebook Papers of CLEF 2011 LABs and Workshops, Amsterdam, The Netherlands.

Antonio , and Manuel. (2010). *CoReMo System (Contextual Reference Monotony)* Paper presented at the the Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN) 2010 Workshop Proceedings of Padua.

Antonio, S. Hong Va Leong, and Rynson, W.H.L. (1997). *CHECK: a document plagiarism detection system.* Paper presented at the Proceedings of the 1997 ACM symposium on Applied computing.

Baker, C.F. Fillmore, C.J., and Lowe, J.B. (1998). *The Berkeley FrameNet Project.* Paper presented at the Proceedings of the 17th international conference on Computational linguistics, V.1 (3)，53-60 .

Balaguer, E.V. (2009). *Putting ourselves in sme's shoes: Automatic detection of plagiarism by the wcopyfind tool*. Proc. SEPLN'09, 34-35.

Bao, J.P., and Malcolm, J. (2006). *Text similarity in academic conference papers*.

Bao, J.P. Shen, J.Y. Liu, X.D., and Song, Q.B. (2003). *A Survey on. Natural Language Text Copy Detection*. Journal of. Software, 14(10), 1753-1760.

Barnickel, T. Weston, J. Collobert, R. Mewes, H., and Stumpflen, V. (2009). *Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts*. International Journal of Information Technology & Decision Making, 4(7), e6393.

Barrón-Cedeño, A.a.P.R. (2009). *On Automatic Plagiarism Detection Based on n-Grams Comparison.* Paper presented at the Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval.

Basile, C. Benedetto, D. Caglioti, E. Cristadoro, G., and Esposti, M.D. (2009). *A plagiarism detection procedure in three steps: Selection, Matches and "Squares"*. Donostia, Spain, 19-23.

Benno Stein Moshe Koppel, and Efstathios Stamatatos. (2007). *Plagiarism analysis, authorship identification, and near-duplicate detection PAN'07*. *SIGIR Forum, 41*(2), 68-71.

Berry, A., and Sigayret, A. (2004). *Representing a concept lattice by a graph*. Discrete Applied Mathematics, 144(1), 27-42.

Bezdek, J.C., and Pal, S.K. (1992). *Fuzzy models for pattern recognition*.

Bloomfield, L. (2008). *'WCopyFind Software.* http://plagiarism. phys. virginia. edu/Wsoftware. html (accessed June 1, 2008).

Breiman, L. Friedman, J. Olshen, R., and Stone, C. (1984). *Classification and Regression Trees:* Wadsworth. I984.

Broder, A.Z. (1997, 11-13 Jun 1997). *On the resemblance and containment of documents.* Paper presented at the Compression and Complexity of Sequences 1997. Proceedings.

Buckley, C. Salton, G. Allan, J., and Singhal, A. (1995). *Automatic query expansion using SMART: TREC 3*. Nist special publication, 69-69.

Buell, D.A. (1982). *An analysis of some fuzzy subset applications to information retrieval systems*. Fuzzy Sets and Systems, 7(1), 35-42.

Bunke, H. (1999). *Error correcting graph matching: On the influence of the underlying cost function*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 21(9), 917-922.

Bunke, H. Jiang, X., and Kandel, A. (2000). *On the minimum common supergraph of two graphs*. Computing, 65(1), 13-25.

Bunke, H., and Shearer, K. (1998). *A graph distance metric based on the maximal common subgraph*. Pattern Recognition Letters, 19(3), 255-259.

Byung-Ryul, A. Heon, K., and Moon-Hyun, K. (2006, Nov. 2006). *An Application of Detecting Plagiarism using Dynamic Incremental Comparison Method.* Paper presented at the International Conference on Computational Intelligence and Security, 2006.

Carreras, X., and Màrquez, L. (2005). Introduction to the CoNLL-2005 shared task: Semantic role labeling.

Chong, M. Specia, L., and Mitkov, R. (2010). *Using Natural Language Processing for Automatic Detection of Plagiarism*. Proceedings of the 4th International Plagiarism.

Chow, and Rahman. (2009). *Multilayer SOM with tree-structured data for efficient document retrieval and plagiarism detection*. Trans. Neur. Netw., 20(9), 1385-1402.

Chow Kent, and Salim, N. (2010). *Features Based Text Similarity Detection*. Journal of Computing, 2(1), 53-57.

Chow Kent, and Salim, N. (2010). *Web Based Cross Language Plagiarism Detection*. Second International Conference on Computational Intelligence, Modelling and Simulation, 199-204.

Claudia Leacock George A. Miller, and Martin Chodorow. (1998). *Using corpus statistics and WordNet relations for sense identification*. Comput. Linguist., 24(1), 147-165.

Clough, P. (2000). *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.

Clough, P. Gaizauskas, R., and Piao, S. (2002). *Building and annotating a corpus for the study of journalistic text reuse*.

Clough, P., and Stevenson, M. (2009). *Creating a Corpus of Plagiarised Academic Texts*.

Coffman, T. Greenblatt, S., and Marcus, S. (2004). *Graph-based technologies for intelligence analysis*. Communications of the ACM, 47(3), 45-47.

Collobert, R., and Weston, J. (2007). *Fast semantic extraction using a novel neural network architecture*. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (June 2007), pp. 560-567.

Cook, D.J., and Holder, L.B. (1994). *Substructure discovery using minimum description length and background knowledge*. Arxiv preprint cs/9402102.

Dahlmeier, D., and Ng, H.T. (2010). *Domain adaptation for semantic role labeling in the biomedical domain*. Bioinformatics, 26(8), 1098-1104.

Daniel Gildea, and Daniel Jurafsky. (2002). *Automatic labeling of semantic roles*. Comput. Linguist., 28(3), 245-288.

Daniel R. White, and Mike S. Joy. (2004). *Sentence-based natural language plagiarism detection*. J. Educ. Resour. Comput., 4(4), 2.

Devi, S.L. Rao, P.R.K. Ram, V.S., and Akilandeswari, A. (2010). *External Plagiarism Detection.* Paper presented at the CLEF (Notebook Papers/LABs/Workshops)

Dinu, L.P., and Popescu, M. (2009). *Ordinal measures in authorship identification*. Proc. SEPLN'09, 62-66.

Djoko, S. Cook, D.J., and Holder, L.B. (1997). *An empirical study of domain knowledge and its benefits to substructure discovery*. Knowledge and Data Engineering, IEEE Transactions on, 9(4), 575-586.

Dowty, D. (1991). *Thematic proto-roles and argument selection*. Language, 67(3), 547-619.

Du Zou Wei-jiang Long, and Zhang Ling. (2010). *A Cluster-Based Plagiarism Detection Method*. [Lab Report for PAN at CLEF 2010]. CLEF (Notebook Papers/LABs/Workshops)

Efstathios Stamatatos. (2008). *Author identification: Using text sampling to handle the class imbalance problem*. Inf. Process. Manage., 44(2), 790-799.

Efstathios Stamatatos. (2009). *A survey of modern authorship attribution methods*. J. Am. Soc. Inf. Sci. Technol., 60(3), 538-556.

Elhadi, M., and Al-Tobi, A. (2008, 13-16 Nov. 2008). *Use of text syntactical structures in detection of document duplicates.* Paper presented at the Third International Conference on.Digital Information Management, 2008. ICDIM 2008.

Elhadi, M., and Al-Tobi, A. (2009). *Duplicate Detection in Documents and WebPages Using Improved Longest Common Subsequence and Documents Syntactical Structures.* Paper presented at the Proceedings of the 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology.

Erkan, G., and Radev, D.R. (2004). *LexRank: Graph-based lexical centrality as salience in text summarization*. J. Artif. Intell. Res. (JAIR), 22, 457-479.

Fellbaum, C. (1998). *WordNet: An electronic database:* MIT Press, Cambridge, MA.

Fernández, M.L., and Valiente, G. (2001). *A graph distance metric combining maximum common subgraph and minimum common supergraph*. Pattern Recognition Letters, 22(6-7), 753-758.

Fillmore, C.J. (1968). *The case for case.* In Emmon Bach and Robert T. Universals in Linguistic Theory. Holt, Rinehart, and Winston, New York, 1–210.

Frakes, W.B., and Baeza-Yates, R. (1992). *Information Retrieval: Data Structures and Algorithm.*

Gader, P. Forester, B. Ganzberger, M. Gillies, A. Mitchell, B. Whalen, M., and Yocum, T. (1991). *Recognition of handwritten digits using template and model matching*. Pattern recognition, 24(5), 421-431.

Gallagher, B. (2006a). *Matching structure and semantics: A survey on graph-based* pattern matching. AAAI FS, 6, 45-53.

Gallagher, B. (2006b). *The state of the art in graph-based pattern matching*: United States. Dept. of Energy.

Gipp, B. and J. Beel (2010). *Citation based plagiarism detection: a new approach to identify plagiarized work language independently.* Proceedings of the 21st ACM conference on Hypertext and hypermedia, ACM.

Gipp, B. and N. Meuschke (2011). *Citation pattern matching algorithms for citation-based plagiarism detection: greedy citation tiling, citation chunking and longest common citation sequence.* Proceedings of the 11th ACM Symposium on Document Engineering (DocEng2011).

Gipp, B., N. Meuschke, et al. (2011). *Comparative evaluation of text-and citation-based plagiarism detection approaches using guttenplag.* Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries, ACM.

Greenblatt, S. Marcus, S., and Darr, T. (2005). *Tmods-integrated fusion dashboard-applying fusion of fusion systems to counter-terrorism*. Proc. International Conference on Intelligence Analysis.

Grman, J., and Ravas, R. (2011). *Improved implementation for finding text similarities in large collections of data.* Paper presented at the CLEF (Notebook Papers/LABs/Workshops)

Grozea, C. Gehl, C., and Popescu, M. (2009). *ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection*. Donostia, Spain, 10-18.

Grozea, C. Gehl, C., and Popescu, M. (2010). *ENCOPLOT : Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection*. in Proc. SEPLN, Donostia, Spain.

Grozea, C., and Popescu, M. (2011). The Encoplot Similarity Measure for Automatic Detection of Plagiarism. Paper presented at the Notebook for PAN at CLEF 2011. In Notebook Papers of CLEF 2011 LABs and Workshops.

Guarino, N. Masolo, C., and Vetere, G. (1999). *Ontoseek: Content-based access to the web*. Intelligent Systems and Their Applications, IEEE, 14(3), 70-80.

Heintze, N. (1996). *Scalable document fingerprinting*. USENIX Workshop on Electronic Commerce, 191-200.

Heon, K. Yang-koo, K. Pyung-Jin, K., and Moon-Hyun, K. (2005, 24-26 May 2005). *An application of DICOM architecture for detecting plagiarism in natural language.* Paper presented at Ninth International Conference on the Computer Supported Cooperative Work in Design, 2005.

Hermann, M. Frank, K., and Bilal, Z. (2006). *Plagiarism - A Survey*. Journal of Universal Computer Science, 12(8), 1050-1084

Höppner, F. (1999). *Fuzzy cluster analysis: methods for classification, data analysis, and image recognition*: Wiley.

Hull, D.A. (1996). *Stemming algorithms: A case study for detailed evaluation*. Journal of the American Society for Information Science, 47(1), 70-84.

Ibrahim, A.M. (2004). *Fuzzy logic for embedded systems applications*: Newnes.

Jaccard, and Paul. (1901). *Étude comparative de la distribution florale dans une portion des Alpes et des Jura*. Bulletin de la Société Vaudoise des Sciences Naturelles, 37, 547–579.

Jang, J.S.R. (1993). *ANFIS: Adaptive-network-based fuzzy inference system*. *Systems,* Man and Cybernetics, IEEE Transactions on 23(3): 665-685.Systems, Man and Cybernetics, IEEE Transactions on, 23(3), 665-685.

Jeh, G., and Widom, J. (2002). *SimRank: a measure of structural-context similarity*. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.

Johansson, R., and Nugues, P. (2008). *The effect of syntactic representation on semantic role labeling.* Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, Association for Computational Linguistics.

Jonker, J. Franses, P.H., and Piersma, N. (2002). *Evaluating direct marketing campaigns; recent findings and future research topics* (ERIM Reports ERS): Erasmus Universiteit Rotterdam.

Joy, A., and Luck, M.M. (1999). *Plagiarism in Programming Assignments*. IEEE Transactions on Education 42(1), 129-133.

Kang, N. Gelbukh, A., and Han, S.-Y. (2006). *PPChecker: Plagiarism pattern checker in document copy detection*. *LNCS LNAI, 4188,* 661–667.

Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms*: Wiley-IEEE Press.

Kantardzie, M., and SRIVASTAVA, A.N. (2005). *Data Mining: concepts, models, methods, and algorithms*. Journal of Computing and Information Science in Engineering, 5(4), 265-395.

Karin Kipper Hoa Trang Dang, and Martha Palmer. (2000). *Class-Based Construction of a Verb Lexicon.* Paper presented at the Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence.

Kasprzak, J., and Brandejs, M. (2009). *Finding Plagiarism by Evaluating Document Similarities*. Donostia, Spain, 24-28.

Kass, G.V. (1980). *An Exploratory Technique for Investigating Large Quantities of Categorical Data*. Journal of the Royal Statistical Society. Series C (Applied Statistics), 29(2), 119-127.

Kestemont, M. Luyckx, K., and Daelemans, W. (2011). *Intrinsic plagiarism detection using character trigram distance scores.* Notebook for PAN at CLEF 2011.

Khoury, R. Karray, F. Sun, Y. Kamel, M., and Basir, O. (2007). *Semantic understanding of general linguistic items by means of fuzzy set theory*. Fuzzy Systems, IEEE Transactions on, 15(5), 757-771.

Kienreich, W. Granitzer, M. Sabol, V., and Klieber, W. (2006, 0-0 0). *Plagiarism Detection in Large Sets of Press Agency News Articles.* Paper presented at the 17th International Workshop on.Database and Expert Systems Applications, 2006. DEXA '06.

Kim, D.H. Yun, I.D., and Lee, S.U. (2004). *A comparative study on attributed relational gra matching algorithms for perceptual 3-D shape descriptor in MPEG-7*.

Kleinberg, J.M. (1999). *Authoritative sources in a hyperlinked environment*. Journal of the ACM (JACM), 46(5), 604-632.

Klir, G., and Yuan, B. (1995). *Fuzzy Sets and Fuzzy Logic:Theory and Applications*. Prentice Hall Upper Saddle River, NJ.

Koroutchev, K., and Cebrian, M. (2006). *Detecting translations of the same text and data with common source*. Statistical Mechanics: Theory and Experiment, 2006(10), 10009-10009.

Kreinovich, V. Quintana, C., and Reznik, L. (1992). *Gaussian membership functions are most adequate in representing uncertainty in measurements*. Proceedings of NAFIPS.

Lancaster, T. (2003). *Effective and efficient plagiarism detection.* South Bank University.

Lennon, M. Pierce, D.S. Tarry, B.D., and Willett, P. (1981). *An evaluation of some conation algorithms for information retrieval*. Journal of Information Science, 3(4), 177-183.

Levenshtein, V.I. (1966). *Binary codes capable of correcting deletions, insertions, and reversals*. Soviet Physics Doklady, 10(8), 707-710.

Lin, and Chin-Yew. (2004). *ROUGE: A Package For Automatic Evaluation Of Summaries*. Workshop On Text Summarization Branches Out.

Lin, D. (1998). *An information-theoretic definition of similarity*. In Proceedings of the Fifteenth International Conference on Machine Learning (ICML), 296–304.

Lönneker-Rodman, B., and Baker, C.F. (2009). *The FrameNet model and its applications*. Natural Language Engineering, 15(3), 414-453.

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

Lyon, C. Barrett, R., and Malcolm, J. (2006). *Plagiarism is easy, but also easy to detect. Plagiary:* Cross-Disciplinary Studies in Plagiarism, Fabrication, and Falsification, 1.

Lyon, C. Malcolm, J.A., and Dickerson, R.G. (2001). *Detecting short passages of similar text in large document collections*. Empirical Methods in Natural Language Processing.

M. K. M. Rahman, and Tommy W. S. Chow. (2010). *Content-based hierarchical document organization using multi-layer hybrid network and tree-structured features*. Expert Syst. Appl., 37(4), 2874-2881.

Malcolm, J., and Lane, P.C.R. (2009). *Tackling the PAN'09 External Plagiarism Detection Corpus with a Desktop Plaigiarism Detector*. pp. 29-33.

Mamdani, E.H. (1974). *Application of fuzzy algorithms for control of simple dynamic plant*. *Electrical Engineers,* Proceedings of the Institution of, 121(12), 1585-1588.

Mamdani, E.H., and Assilian, S. (1975). *An experiment in linguistic synthesis with a fuzzy logic controller*. International Journal of Man-Machine Studies, 7(1), 1-13.

Markov, A., and Last, M. (2005). *Efficient graph-based representation of web documents*. MGTS 2005, 51.

Markov, A. Last, M., and Kandel, A. (2006). *Model-based classification of web documents represented by graphs.* WebKDD: Workshop on Web Mining and Web Usage Analysis.

Màrquez, L. Carreras, X. Litkowski, K.C., and Stevenson, S. (2008). *Semantic role labeling: an introduction to the special issue*. Computational linguistics, 34(2), 145-159.

Martha Palmer Daniel Gildea, and Paul Kingsbury. (2005). *The Proposition Bank: An Annotated Corpus of Semantic Roles*. Comput. Linguist., 31(1), 71-106.

MathWorks. (2012). *Fuzzy Logic Toolbox™ User's Guide* http://www.mathworks.com/help/pdf_doc/fuzzy/fuzzy.pdf.

Maurer, H. Kappe, F., and Zaka, B. (2006). *Plagiarism-a survey*. Journal of Universal Computer Science, 12(8), 1050-1084.

McCabe, D. (2005). *Research Report of the Center for Academic Integrity*.

McKay, B. (1990). *nauty user's guide (version 1.5),* Computer Science Department, Australian National University: Tech. Rep. TR-CS-90-02.

Merel van Diepen, and Philip Hans Franses. (2006). *Evaluating chi-squared automatic interaction detection*. Inf. Syst., 31(8), 814-831.

Messmer, B., and Bunke, H. (1996). *Subgraph isomorphism detection in polynomial time on preprocessed model graphs*. Recent Developments in Computer Vision, 373-382.

Meyer zu Eissen, S. Stein, B., and Kulig, M. (2007). *Plagiarism Detection Without Reference Collections. Advances in Data Analysis*. In R. Decker and H.J. Lenz (Eds.), (pp. 359-366): Springer Berlin Heidelberg.

Micol, D. Ferrández, Ó. Llopis, F., and Muñoz, R. (2010). *A Textual-Based Similarity Approach for Efficient and Scalable External Plagiarism Analysis.* Paper presented at the CLEF (Notebook Papers/LABs/Workshops).

Midhun, M. Shine, N.D., and Pramod, K.V. (2011). *A Novel Approach for Near-Duplicate Detection of Web Pages using TDW Matrix*. International Journal of Computer Applications, 19(7), 16-21.

Mikheev, A. (2000). *Document centered approach to text normalization*. In Proceedings of SIGIR, 136-143.

Miller, G., and Fellbaum, C. (1998). *Wordnet: An electronic lexical database*. May, 15, 02142-01493.

Minsky, M. (1974). *A framework for representing knowledge*. The Psychology of Computer Vision, McGraw-Hill: 211-277.

Mogharreban, N., and Dilalla, L.F. (2006, 3-6 June 2006). *Comparison of Defuzzification Techniques for Analysis of Non-interval Data.* Paper presented at the Fuzzy Information Processing Society, 2006. NAFIPS 2006. Annual meeting of the North American.

Mohammed Salem Binwahlan Naomie Salim, and Ladda Suanmali. (2010). *Fuzzy swarm diversity hybrid model for text summarization*. Inf. Process. Manage., 46(5), 571-588.

Monostori, K. Zaslavsky, A., and Schmidt, H. (2000). *Document overlap detection system for distributed digital libraries*. Proceedings of the fifth ACM conference on Digital libraries, ACM.

Morante, R. Asch, V.V., and Bosch, A.v.d. (2009). *Joint memory-based learning of syntactic and semantic dependencies in multiple languages.* Paper presented at the Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task.

Mozgovoy, M. (2007). *Enhancing computer-aided plagiarism detection*: University of Joensuu.

Mozgovoy, M. Fredriksson, K. White, D. Joy, M., and Sutinen, E. (2005). *Fast plagiarism detection system.* String Processing and Information Retrieval, Springer.

Muhr, M. Kern, R. Zechner, M., and Granitzer, M. (2010). *External and Intrinsic Plagiarism Detection using a Cross-Lingual Retrieval and Segmentation System*. Lab Report for PAN.

Myers, R. Wison, R., and Hancock, E.R. (2000). *Bayesian graph edit distance*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22(6), 628-635.

Oberreuter, G. L'Huillier, G. Ríos, S.A., and Velásquez, J.D. (2010). *FASTDOCODE: Finding Approximated Segments of N-Grams for Document Copy Detection*. Lab Report for PAN at CLEF(2010).

Office of Academic Appeals & Regulation of the published procedures. (2011). *Cheating, plagiarism and fraudulent or fabricated coursework.* (available online at http://www.leeds.ac.uk/AAandR/cpff.htm - accessed 28th Nov 2011).

Osen, J. (1997). *The cream of other men's wit: Plagiarism and misappropriation in cyberspace*. Computer Fraud & Security, 1997(11), 13-19.

Pablo Suárez José Carlos González, and Julio Villena-Román. (2010). *in proceedings of the Uncovering Plagiarism Authorship and Social Software Misuse.* Lab Report for PAN at CLEF 2010. CLEF (Notebook Papers/LABs/Workshops).

Page, L. Brin, S. Motwani, R., and Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*. Technical Report, Stanford University Database Group, 1999.

Palkovskii, Y. Belov, A., and Muzyka, I. (2011, 19-22 September). *Using WordNet-based semantic similarity measurement in External Plagiarism Detection.* Paper presented at the CLEF (Notebook apers/LABs/Workshops), Amsterdam, The Netherlands.

Parth, G. Sameeer, R., and Majumdar, P. (2010). *External Plagiarism Detection: N-Gram Approach using Named Entity Recognizer*. Paper presented at the CLEF (Notebook Papers/LABs/Workshops)

Payne, T.E. (1997). *Describing morphosyntax: A guide for field linguists*: Cambridge Univ Pr.

Perreault Jr, W.D., and Barksdale Jr, H.C. (1980). *A model-free approach for analysis of complex contingency data in survey research*. Journal of Marketing Research, 503-515.

Potthast, M. Barrón-Cedeño, A. Eiselt, A. Stein, B., and Rosso, P. (2010). *Overview of the 2nd international competition on plagiarism detection*. Notebook Papers of CLEF, 10.

Potthast, M. Barrón-Cedeño, A. Eiselt, A. Stein, B., and Rosso, P. (2011). *Overview of the 3nd international competition on plagiarism detection*. Notebook Papers of CLEF 11, 10.

Potthast, M. Stein, B. Eiselt, A. Barrón-Cedeño, A., and Rosso, P. (2009). *Overview of the 1st International Competition on Plagiarism Detection.* Paper presented at the PAN-09 3rd Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse and 1st International Competition on Plagiarism Detection.

Pradhan Sameer, S. Wayne, H. Ward, K.H. James, H.M., and Dan Jurafsky. (2004). *Shallow semantic parsing using support vector machines*. In Proceedings of NAACL-HLT 2004, 233–240.

Prechelt, L. Malpohl, G., and Philippsen, M. (2000). JPlag: Finding plagiarism among a set of programs: Technical Report 2000-1, Fakultat fur Informatik, Universitat Karlsruhe, D-76128 Karlsruhe, Germany.

Punyakanok, V. Roth, D., and Yih, W. (2008). *The importance of syntactic parsing and inference in semantic role labeling*. Computational linguistics, 34(2), 257-287.

Rahman, M.K.M. Wang, P.Y. Tommy , W.S.C., and Sitao, W. (2007). *A flexible multi-layer self-organizing map for generic processing of tree-structured data*. Pattern Recogn., 40(5), 1406-1424.

Rao, K. (2008). *Plagiarism, a scourge*. Current Science Bangalore, 94(5), 581.

Rijsbergen, C., and Van, J. (1979). *A New Theoretical Framework for Information Retrieval.*

Robles-Kelly, A., and Hancock, E. (2003). *Edit distance from graph spectra.* Paper presented at the Ninth IEEE International Conference on Computer Vision, 2003.

Robles-Kelly, A., and Hancock, E.R. (2005). *Graph edit distance from spectral seriation*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27(3), 365-378.

Roig, M. (2009). *Avoiding plagiarism, self-plagiarism, and other questionable writing practices: A guide to ethical writing:* United States Department of Health & Human Services. Office of Research Integrity.

Ryu, C.-K. Kim, H.-J., and Cho, H.-G. (2009). *A detecting and tracing algorithm for unauthorized internet-news plagiarism using spatio-temporal document evolution model.* Paper presented at the Proceedings of the 2009 ACM symposium on Applied Computing.

Samuelson, P. (1994). *Self-plagiarism or fair use*. Communications of the ACM, 37(8), 21-25.

Schenker, A. (2005). *Graph-theoretic techniques for web content mining*. World Scientific Publishing Company Incorporated (Vol. 62).

Seaward, L., and Matwin, S. (2009). *Intrinsic plagiarism detection using complexity analysis*. PAN, 9, 56-61.

Sergey Brin James Davis ctor Garc, and Molina. (1995a). *Copy detection mechanisms for digital documents*. SIGMOD Rec., 24(2), 398-409.

Sergey Brin James Davis ctor Garc, and Molina. (1995b). *Copy detection mechanisms for digital documents.* Paper presented at the Proceedings of the 1995 ACM SIGMOD international conference on Management of data.

Setnes, M. Babuska, R. Kaymak, U., and van Nauta Lemke, H.R. (1998). *Similarity measures in fuzzy rule base simplification*. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 28(3), 376-386.

Shapiro, L., and Haralick, R. (1981). *Structural descriptions and inexact matching*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 3, 504-519.

Shasha, D. Wang, J.T.L., and Giugno, R. (2002). *Algorithmics and applications of tree and graph searching.* Paper presented at the Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems.

Shcherbinin, V., and Butakov, S. (2009). *Using Microsoft SQL server platform for plagiarism detection. SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09).*

Shivakumar, N., and Garcia-Molina, H. (1995). *SCAM: A copy detection mechanism for digital documents. ACM SIGMOD Record, ACM.*

Siler, W. Buckley, J.J., and Wiley, J. (2005). *Fuzzy expert systems and fuzzy reasoning*: Wiley Online Library.

Spracklin, L. Inkpen, D., and Nayak, A. (2008). *Using the Complexity of the Distribution of Lexical Elements as a Feature in Authorship Attribution.* LREC 2008 Proceedings, Marrakech, Morocco.

SPSS, S.P. (2011). *Service Solutions*. SPSS Clementine.[Software]. SPSS.

Stamatatos, E. (2009). *Intrinsic plagiarism detection using character n-gram profiles*. Paper presented at the Proceedings of the 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse.

Stefan Gruner, and Stuart Naven. (2005). *Tool support for plagiarism detection in text documents.* Paper presented at the Proceedings of the 2005 ACM symposium on Applied computing.

Stein, B. Lipka, N., and Prettenhofer, P. (2011). *Intrinsic plagiarism analysis*. Language Resources and Evaluation, 45(1), 63-82.

Suanmali, L. Binwahlan, M.S., and Salim, N. (2009a). *Sentence Features Fusion for Text Summarization Using Fuzzy Logic.* Paper presented at the Ninth International Conference on Hybrid Intelligent Systems.

Suanmali, L. Binwahlan, M.S., and Salim, N. (2009b). *Sentence features fusion for text summarization using fuzzy logic*. Hybrid Intelligent Systems, 2009. HIS'09. Ninth International Conference on, IEEE.

Suanmali, L. Salim, N., and Binwahlan, M.S. (2009c). *Automatic Text Summarization Using Feature-Based Fuzzy Extraction*. Jurnal Teknologi Maklumat, 2(1), 105-155.

Sugeno, M., and Takagi, T. (1983). *Multi-dimensional fuzzy reasoning*. Fuzzy Sets and Systems, 9(1-3), 313-325.

Sulema Torres, and Alexander Gelbukh. (2009). *Comparing Similarity Measures for Original WSD Lesk Algorithm*. Advances in Computer Science and Application, 43, 155-166.

Surdeanu, M. Harabagiu, S. Williams, J., and Aarseth, P. (2003). *Using predicate-argument structures for information extraction.* Paper presented at the Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1.

Takagi, T., and Sugeno, M. (1985). *Fuzzy identification of system and its applications to modelling and control*. IEEE Trans. Syst., Man, and Cyber, 1, 5.

Ting, Y. Lu, W. Chen, C., and Wang, G. (2008). *A fuzzy reasoning design for fault detection and diagnosis of a computer-controlled system*. Engineering applications of artificial intelligence, 21(2), 157-170.

Tomasic, A., and Garcia-Molina, H. (1993). *Query processing and inverted indices in shared: nothing text document information retrieval systems*. The VLDB Journal, 2(3), 243-276.

Toshinori Munakata, and Yashvant Jani. (1994). *Fuzzy systems: an overview*. *Commun.* ACM, 37(3), 68-76.

Tsai, W.H., and Fu, K.S. (1979). *Error-correcting isomorphisms of attributed relational graphs for pattern analysis*. Systems, Man and Cybernetics, IEEE Transactions on, 9(12), 757-768.

Ullmann, J.R. (1976). *An algorithm for subgraph isomorphism*. Journal of the ACM (JACM), 23(1), 31-42.

Vania, C., and Adriani, M. (2010). *Automatic External Plagiarism Detection Using Passage Similarities.* Paper presented at the CLEF (Notebook Papers/LABs/Workshops)

Vieira, S.M. Sousa, J.M.C., and Kaymak, U. (2012). *Fuzzy criteria for feature selection*. [doi: 10.1016/j.fss.2011.09.009]. Fuzzy Sets and Systems, 189(1), 1-18.

Vikramjit Mitra Chia-Jiu Wang, and Satarupa Banerjee. (2007). *Text classification: A least square support vector machine approach*. Applied Soft Computing, 7(3), 908-914.

Wallis, W.D. Shoubridge, P. Kraetz, M., and Ray, D. (2001). *Graph distances using graph union*. Pattern Recognition Letters, 22(6-7), 701-704.

Wang, L.X. (1992). *Fuzzy systems are universal approximators*. Computers, IEEE Transactions on 43(11): 1329-1333.

Washio, T., and Motoda, H. (2003). *State of the art of graph-based data mining*. Acm Sigkdd Explorations Newsletter, 5(1), 59-68.

Watson, S.R. Weiss, J.J., and Donnell, M.L. (1979). *Fuzzy decision analysis*. Systems, Man and Cybernetics, IEEE Transactions on, 9(1), 1-9.

Wong, S.V., and Hamouda, A.M.S. (2000). *Optimization of fuzzy rules design using genetic algorithm*. Adv. Eng. Softw., 31(4), 251-262.

Xue Nianwen, and Martha Palmer. (2004). *Calibrating features for semantic role labeling*. In Proceedings of EMNLP 2004, 88–94.

Yerra, R., and Ng, Y.-K. (2005). *A Sentence-Based Copy Detection Approach for Web Documents*. Fuzzy Systems and Knowledge Discovery, 3613, 557-570.

Yih-Jeng, L. Ming-Shing, Y., and Chin-Yu, L. (2008, 26-28 Nov. 2008). *Using Chi-Square Automatic Interaction Detector to Solve the Polysemy Problems in a Chinese to Taiwanese TTS System.* Paper presented at the Eighth International Conference on.Intelligent Systems Design and Applications, 2008. ISDA '08.

Yoo, I. Hu, X., and Song, I.Y. (2006). *Integration of semantic-based bipartite graph representation and mutual refinement strategy for biomedical literature clustering.* Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.

Zadeh, L.A. (1965). *Fuzzy sets*. *Information and Control 8* (3), 338-353.

Zdenek Ceska Michal Toman, and Karel Jezek. (2008). *Multilingual Plagiarism Detection.* Paper presented at the Proceedings of the 13th international conference on Artificial Intelligence: Methodology, Systems, and Applications.

Zechner, M. Muhr, M. Kern, R., and Granitzer, M. (2009). *External and intrinsic plagiarism detection using vector space models*. in Proc. SEPLN, Donostia, Spain.

Zhang, L. Zhuang, Y., and Yuan, Z. (2007). *A program plagiarism detection model based on information distance and clustering*. The 2007 International Conference on Intelligent Pervasive Computing, 2007.

Zhang, X. (2009). *Exploiting external/domain knowledge to enhance traditional text mining using graph-based methods.* Drexel University.

Zini, M. Fabbri, M. Moneglia, M., and Panunzi, A. (2006, 13-15 Dec. 2006). *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. AXMEDIS '06.