DOCUMENT PLAGIARISM DETECTION ALGORITHM USING SEMANTIC NETWORKS

AHMED JABR AHMED MUFTAH

A project report submitted in partial fulfillment of the requirements for the award of the degree of Master of Science (Computer Science)

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

NOVEMBER 2009

Dedicated to my father, who taught me that anything is possible to achieve; the only limitations in our lives are those that we impose on ourselves.

ACKNOWLEDGEMENT

Most of my gratitude goes to my supervisor Assoc. Prof. Dr. Naomie. Her patience and considerate nature made her accessible whenever I needed her assistance. I indeed thank her for showing me how to identify interesting problems and how a research can be started and finished correctly.

I acknowledge that my UTM colleagues are the greatest. My especial thank to Omar and Mohammed Hakami for their unrelenting encouragement during project-1. Also many thanks to Ali Alfaris, Amjad Esmaiel, Murad Rassam, Yassir, and Falah for helping me retain some sanity to get this work done.

Last but not less, I thank my beloved brothers Esam, Nasser, Hesham and Hussam for their ultimate support during the course of my study especially in my final semester. Hey guys thanks for everything.

ABSTRACT

The vast increase of available documents in the World Wide Web (WWW) and the ease access to these documents has lead to a serious problem of using other's works without giving credits. Although many methods have been developed to detect some instances of plagiarism such as changing the structure of sentences or when slightly replacing words by their synonyms, it is often hard to reveal plagiarism when the copied sentences are deliberately modified. This project proposes an algorithm for plagiarism detection over the Web using semantic networks. The corpus of this study contains 610 documents downloaded from the Web, 10 of those were selected to be the source of 20 manually plagiarized documents. The algorithm was compared to N-grams representation and the achieved results show that an appropriate semantic representation of sentences derived from WordNet's relations outperforms N-grams with different similarity measures in detecting the plagiarized sentences. It also show that a proposed method based on extracting named entities and common nouns is in-general capable for retrieving the source documents from the Web using a search engine API when sentences are being moderately plagiarized.

ABSTRAK

Peningkatan keluasan sedia dokumen-dokumen di World Wide Web (WWW) dan kemudahan akses kepada dokumen-dokumen ini telah menyebabkan masalah yang serius dengan menggunakan karya-karya lain tanpa memberikan kredit. Walaupun banyak kaedah telah dibangunkan untuk mengesan beberapa kes plagiarisme seperti menukar struktur kalimat ataupun ketika sedikit menukar kata dengan mereka sinonim, sering sukar untuk mendedahkan plagiarisme ketika menyalin kalimat-kalimat yang sengaja diubahsuai. Projek ini mencadangkan sebuah algoritma untuk mengesan plagiarisme melalui Web menggunakan rangkaian semantik. Korpus kajian ini mengandungi 610 dokumen-download dari Web, 10 daripada mereka yang terpilih untuk menjadi sumber secara manual menjiplak daripada 20 dokumen. Algoritma ini dibandingkan dengan N-gram representasi dan keputusan yang dicapai menunjukkan bahawa representasi semantik yang tepat dari kalimat yang berasal dari hubungan WordNet melebihi N-gram dengan berbagai ukuran persamaan dalam mengesan menjiplak kalimat. Hal ini juga menunjukkan bahawa kaedah yang dicadangkan berdasarkan pada ekstraksi bernama entiti dan kata benda umum adalah jeneral yang mampu untuk mengambil dokumen-dokumen sumber dari Web menggunakan enjin pencari API ketika kalimat-kalimat yang sedang sedang dijiplak.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

The World Wide Web (Web) is the biggest source of information these days. People now can easily search for, access, and browse Web pages to get the information they need, one can imagine how difficult the academic research would be without the Internet and the Web. It is also now easy, and again because the scale and the digital structure of the Web, to use someone else's work illegally.

The problem of plagiarism has its direct association to academia. Maurer et al. [3] defined it as "the unacknowledged used of someone else's work". The most common type is written-text plagiarism in which the plagiarized document is formed by copying some or all parts of the original document(s) possibly with some alterations. Plagiarism is classified into *intra-corpal* and *extra-corpal* with respect to the location of the source document(s) [1]. The former happens when both the copy and source documents are within the same corpus, such as within a collection of students' submissions or within a digital library. While in the latter, the copy and

source documents are not of the same corpus. Here the source documents could be from textbooks or most commonly Web documents. Unless the problem of locating the source documents is solved, it is hard to prove this kind of plagiarism. Identifying Web documents from which copying has occurred is stressful and time consuming for a human inspector given the large number of documents that need to be compared. As the digital structure of Web documents made it easy to plagiarize, fortunately it means that such instances of plagiarism could be traced in an automated manner.

There are two methods to provide an access to a large number of Web documents. The first method is by indexing documents through Web crawling; this has the inherent problems of Web documents that face any Web retrieval system such as bulk size, heterogeneity, and duplication [2], however the system could be tuned for the retrieval purposes, for example if the purpose is to detect plagiarism, the system can be employed to return the most syntactically or semantically similar documents to the query document. The other method, which this project will use, is utilizing general-purposes search engines (such as Google, Yahoo, and Bing) as they provide access services to their systems. The suspected document can be considered as a sequence of queries submitted to the search engine, the result are then compared with the input document.

Intuitively it is required to partition the query document into more primitive units plausible for querying the search engine and for documents comparisons. Sentences are suitable for both cases since they carry ideas and also plagiarism patterns (e.g., insertion, deletion, and/or substitution).

Similarity between sentences (or more generally objects) can be captured numerically using similarity measures such as Jaccard similarity, Overlap similarity, Cosine similarity. These measures are called symmetric functions and widely used

in many Information Retrieval applications. Each measure returns a value indicating the degree of similarity between pairs of objects usually between 0 and 1.

Beside the similarity measures, another aspect is the document (or sentence) representation. There are many representations that have been developed including document fingerprinting [17], bag-of-word model [10], N-grams (consecutive words of length N). Another important representation comes from semantic networks. A semantic network or net "is a graphic notation for representing knowledge in patterns of interconnected nodes and arcs" [50]. Concepts in semantic networks are usually organized in hierarchal structure as illustrated in Figure 1.1

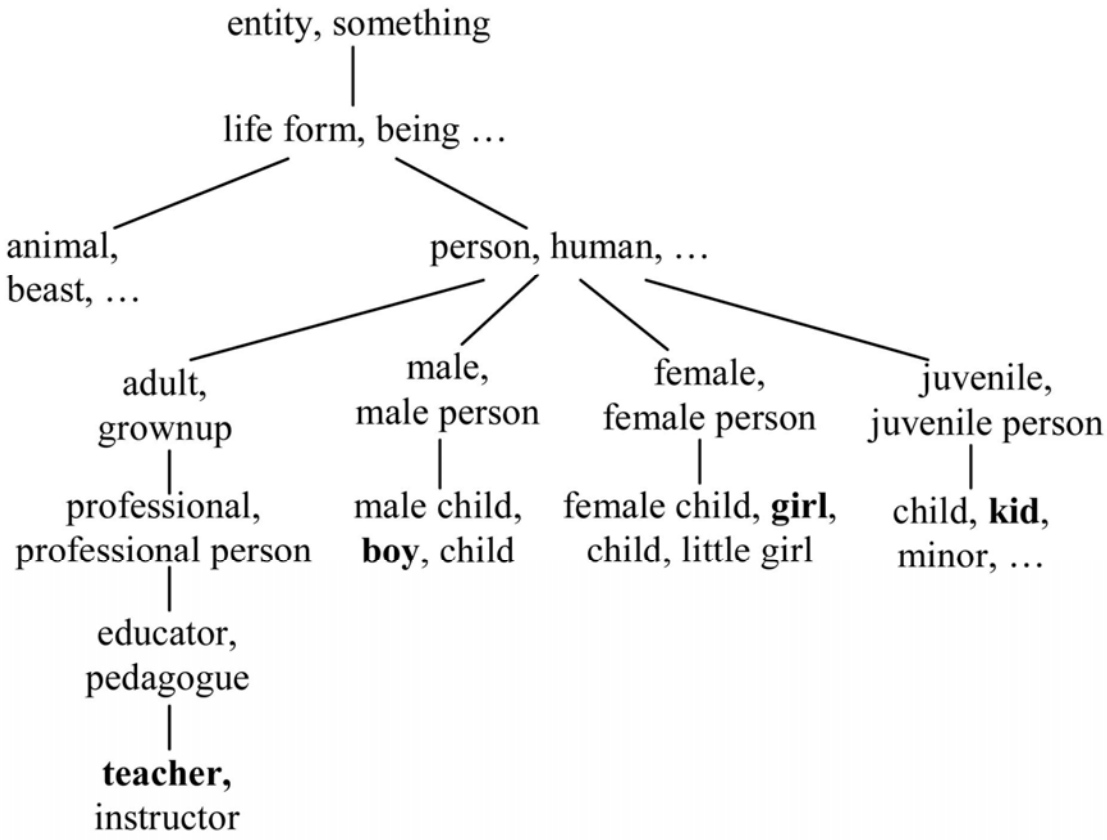

Figure 1.1 Hierarchical semantic knowledge base[57].

Usually words at upper layers of hierarchical semantic nets have more general concepts and less semantic similarity between words than words at lower layers [57].

1.2 Problem Background

In any application that involve measuring the similarity between textual contents there are two important factors that influence the accuracy of plagiarism detection. The first factor is the document representation which essentially captures the characteristics of the document as a preceding step to the comparison stage. These representations include the “Bag-of-Word” model, document Fingerprints, N-grams, probabilistic models. Most of these representations work well in detecting verbatim (word-to-word) plagiarism but have vulnerabilities in detecting complicated plagiarism patterns.

The second factor is the similarity measure that is used to calculate the similarity or dissimilarity between sentences. Considering the plagiarists behavior that usually involves insertions of words deletions and/or substitutions it is necessary to determine which measure is the best for detecting instances of plagiarism.

Retrieving the source documents from the Web using a search engine is another challenge given the fact that some plagiarism patterns are hard to locate in the setting of the Web even for a human inspector.

In this project we investigate the effectiveness of semantic net-based techniques for detecting plagiarized sentences and find out whether the achieved performance is justified comparing to other approaches. Then we determine which technique is the best for retrieving the source documents from the Web.

1.3 Problem Statement

To cater the problems introduced in section 1.2, this project is carried out to answer the following questions:

i- Which N-gram representation is the best for sentence-based plagiarism detection?

ii- Which similarity measure is the best for sentence-based plagiarism detection?

iii- How can semantic networks be used to improve the detection?

1.4 Project Objectives

The main objectives of this project are stated as follows:

i- To compare the effectiveness of different N-gram with different similarity measures in detecting plagiarized documents over the Web.

ii- To find out whether the use of semantic networks can improve the detection of plagiarized documents.

1.5 Project Scope

i- This project will cover plagiarism detection in English scripts.
ii- WordNet [4] is the general semantic network that will be used in this study.
iii- N-grams will be used with three symmetric measures; Cosine, Jaccard, and Dice coefficients.
iv- Porter algorithm [60] will be applied in the stemming process.

1.6 Project Justification

The problem of document plagiarism detection is not new and several methods have been applied to overcome this problem over a small collection of documents or digital libraries, however, the scale of the problem has increased dramatically due to the Web.

It is also widely acceptable that traditional methods for measuring the similarity between documents are vulnerable to fail in some complex plagiarism patterns and hence it is necessary to incorporate semantic-based techniques for more accurate plagiarism detection.

1.7 Report Organization

This report is organized as follows:

Chapter 1 formulates the problem and outlines the framework and main objectives of the project.

Chapter 2 consists of four main parts; the first part introduces some terminologies of document plagiarism detection and briefly outlines some plagiarism detection methods. The second part focuses on semantic networks, in particular WordNet and its semantic relations. The third part is then devoted to document pre-processing and representation techniques and their effect in the applications of plagiarism detection, it also reviews the main approaches for semantic relatedness between concepts. The last part reviews efficient exact set similarity algorithms and discusses how they can be adopted in the case of N-grams.

Chapter 3 illustrates the methodology that will be used to fulfill the objectives of this project.

Chapter 4 presents the experimental results of this project, and finally chapter 5 concludes this research.