

**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 Esmail, 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

CHAPTER	CONTENT	PAGE
	TITLE	i
	DECLARATION	ii
	DEDICATION	iii
	AKNOWLEDGEMENT	iv
	ABSTRACT	v
	ABSTRAK	vi
	TABLE OF CONTENTS	vii
	LIST OF TABLES	x
	LIST OF FIGURES	xii
	LIST OF APPENDICES	xv
1	<b>INTRODUCTION</b>	<b>1</b>
	1.1 Introduction	1
	1.2 Problem Background	4
	1.3 Problem Statement	5
	1.4 Project Objectives	5
	1.5 Project Scope	6
	1.6 Project Justification	6
	1.7 Report Organization	7
2	<b>LITERATURE REVIEW</b>	<b>8</b>
	2.1 Introduction	8

2.2	Document Plagiarism	9
2.3	Plagiarism Detection Methods	10
	2.3.1 Detection based on Stylometry Analysis	11
	2.3.2 Detection based on Documents Comparison	12
	2.3.2.1 Semantic-Based Detection	12
	2.3.2.2 Syntactic-Based Detection	14
2.4	Existing Web-Based Plagiarism Detection Tools	16
2.5	Semantic Networks	20
2.6	Document Preprocessing	23
	2.6.1 Tokenization	24
	2.6.2 Stop-word Removal	24
	2.6.3 Stemming	24
	2.6.4 Document Chunking	25
2.7	Document Representations & Similarity Measures	28
	2.7.1 Semantic Based-Representation	29
	2.7.2 Syntactic Based-Representation	34
	2.7.2.1 Fingerprinting	34
	2.7.2.2 Term Weighting Schemes	37
	2.7.2.3 N-Grams	38
2.8	Algorithms for Approximate Similarity	40
	2.8.1 Signature Scheme Algorithms	40
	2.8.2 Inverted Index-Based Algorithms	47
2.9	Discussion and Summary	51
<b>3</b>	<b>METHODOLOGY</b>	<b>53</b>
3.1	Introduction	53
3.2	Operational Framework	54
	3.2.1 Initial Study and Literature Review	55
	3.2.2 Corpus Preparation	55
	3.2.3 Document Preprocessing	57
	3.2.4 Applying Plagiarism Detection Techniques	58
	3.2.4.1 Semantic Relatedness Approach	58

	3.2.4.2 N-grams Approach	66
	3.2.5 Web Document Retrieval	70
	3.2.6 Implementation	73
	3.2.7 Findings Evaluation	75
<b>4</b>	<b>EXPERIMENTAL RESULTS</b>	<b>78</b>
	4.1 Introduction	78
	4.2 Information about the Corpus	79
	4.3 Sentence-to-Sentence similarity	82
	4.3.1 N-grams Approach	82
	4.3.2 Semantic Relatedness Approach	85
	4.4 Results and Comparisons	97
	4.4.1 Results of Corpus Sentence Retrieval	97
	4.4.2 Results of Web Document Retrieval	102
	4.4.3 Comparison with Existing Tools	111
	4.5 Discussion and Summary	115
<b>5</b>	<b>CONCLUSION</b>	<b>117</b>
	5.1 Introduction	117
	5.2 Achievements and Constraints	118
	5.3 Future work	119
	5.4 Summary	121
	<b>REFERENCES</b>	<b>122</b>
	APPENDICES A-E	128-150

## LIST OF TABLES

TABLE NO.	TITLE	PAGE
2.1	Properties of some existing plagiarism detection tools based on [59]	19
2.2	Some of the relations between concepts in WordNet (N=noun, V=Verb, Adj=adjective, Adv=adverb)	21
2.3	Statistics about WordNet 2.1	22
2.4	Common similarity measures between binary vectors	39
2.5	Common similarity measures between sets	39
2.6	Common factors that influence the performance of Inverted index algorithms	48
2.7	Signature-based versus Inverted index-based algorithms	52
3.1	Integrated libraries in the project and their roles	73
4.1	Number of plagiarized sentences in documents pairs (query-vertical)/(source -horizontal)	80
4.2	Statistics about the corpus and query documents	81
4.3	Statistics about part-of-speech tagging	81
4.4	Part-of-speech tagging of $s_1$ and $s_2$	86

4.5	Shortest path between word pairs in the joint set and $T_2$ (“-1” no path exists, “=” equals, “?” not of the same part of speech)	91
4.6	Subsumer depth between word pairs in the joint set and $T_2$ (-1 no depth exists, = equals, ? not of the same part of speech)	91
4.7	Word-to-word similarity between the joint set and $T_2$	92
4.8	Raw semantic and order vectors for $T_1$	93
4.9	Raw semantic and order vectors for $T_2$	94
4.10	Information contents of word in the joint set	95
4.11	N-grams recall rate in 610 corpus documents with 0.5 cutoff threshold	97
4.12	Recall rate when increasing number of documents with 0.5 cutoff threshold	99
4.13	Precision, Recall, and Harmonic Mean (F-measure) in 110 corpus documents with 0.5 cutoff threshold	100
4.14	Recall rate across similarities in 110 corpus documents	100
4.15	Semantic-R recall rate in 110 corpus documents with Alpha=0.2, Beta=0.45 and 0.8 cutoff threshold	102
4.16	Results of using 3-grams searching with 64 results/query limit	104
4.17	Results of using weighted 3-grams with 64 results/query limit	105
4.18	Results of using selective searching with 64 results/query limit.	108
4.19	Results of using 3-grams searching with 8 results/query limit.	109
4.20	Results of using weighted 3-grams with 8 results/query limit.	110
4.21	Results of using selective searching with 8 results/query limit.	111

## LIST OF FIGURES

FIGURE NO.	TITLE	PAGE
1.1	Hierarchical semantic knowledge base[57]	3
2.1	Taxonomy of plagiarism detection methods	10
2.2	An example HTML report generated by DocCop	16
2.3	A sample report with timeline returned from Plagium	17
2.4	The interface of EVE2 for Web searching	18
2.5	An example of a synset in WordNet	20
2.6	Bipolar adjective structure ( $\rightarrow$ = similarity, $--\rightarrow$ =antonymy)	22
2.7	N-Unit non-overlapped chunking strategy with N=5	26
2.8	N-Unit chunking with K-overlap where N=5 and K=2	26
2.9	An example given by [55] to illustrate the different between most specific subsumers in WordNet.	32
2.10	Framework for signature-based algorithms [7]	41
2.11	Two documents represented as records	41

2.12	Prefix Filter scheme of Figure 2.5 with 80% Overlap similarity threshold.	42
2.13	Two vectors with hamming distance = 4	43
2.14	Two vectors with hamming distance $\leq k$ 4 must agree on one of the $k + 1$ partitions	44
2.15	The two vectors in Figure 2.14 with hamming distance = 8 and agree on one partition	44
2.16	Enumeration scheme for two vectors with hamming distance=3	45
2.17	Formal specification of PartEnum[7]	46
2.18	Formal specification of All-Pairs[6]	49
3.1	Operational Framework	54
3.2	The procedure used in obtaining the semantic attributes between two concepts	63
3.3	The algorithm for semantic relatedness between a pair of sentences	65
3.4	Binary vector representation of a sentence	66
3.5	An inverted index implementation for Cosine similarity[6]	67
3.6	An inverted index implementation for Jaccard Similarity	70
3.7	An inverted index implementation for Dice coefficients	70
3.8	The procedure of evaluating Web document retrieval techniques	72
3.9	Response format from querying the Google API	74
4.1	The inverted index for document $Q$	84
4.2	The hypernym trees for the words: <i>Idea</i> (5 senses), <i>Learning</i> (2 senses), <i>Adaptation</i> (3 senses), <i>Evolution</i> (2 senses) and <i>Biology</i> (3 senses).	89

4.3	The hypernym trees for the words: <i>Idea</i> (5 senses), <i>Learning</i> (2 senses), <i>Adaptation</i> (3 senses), <i>Evolution</i> (2 senses) and <i>Aspect</i> (3 senses)	90
4.4	Recall rate (y-axis) across similarities (x-axis) in 110 corpus documents	101
4.5	Recall rate in one-to-one exact copies	113
4.6	Recall rate in one-to-one plagiarized by synonym replacing	114
4.7	Recall rate in one-to-many plagiarized by synonym replacing	114

**LIST OF APPENDICES**

<b>APPENDIX</b>	<b>TITLE</b>	<b>PAGE</b>
A	Stop-words and their corresponding frequencies in the Brown corpus	128
B	Information about ScienceDirect source documents	129
C	Information about Wikipedia corpus documents	130
D	The PENN TreeBank English POS tag set and their mappings	144
E	Examples of original/plagiarized sentence pairs and the corresponding similarities based on equation 3.6	145

## 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-corp*al and *extra-corp*al 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.

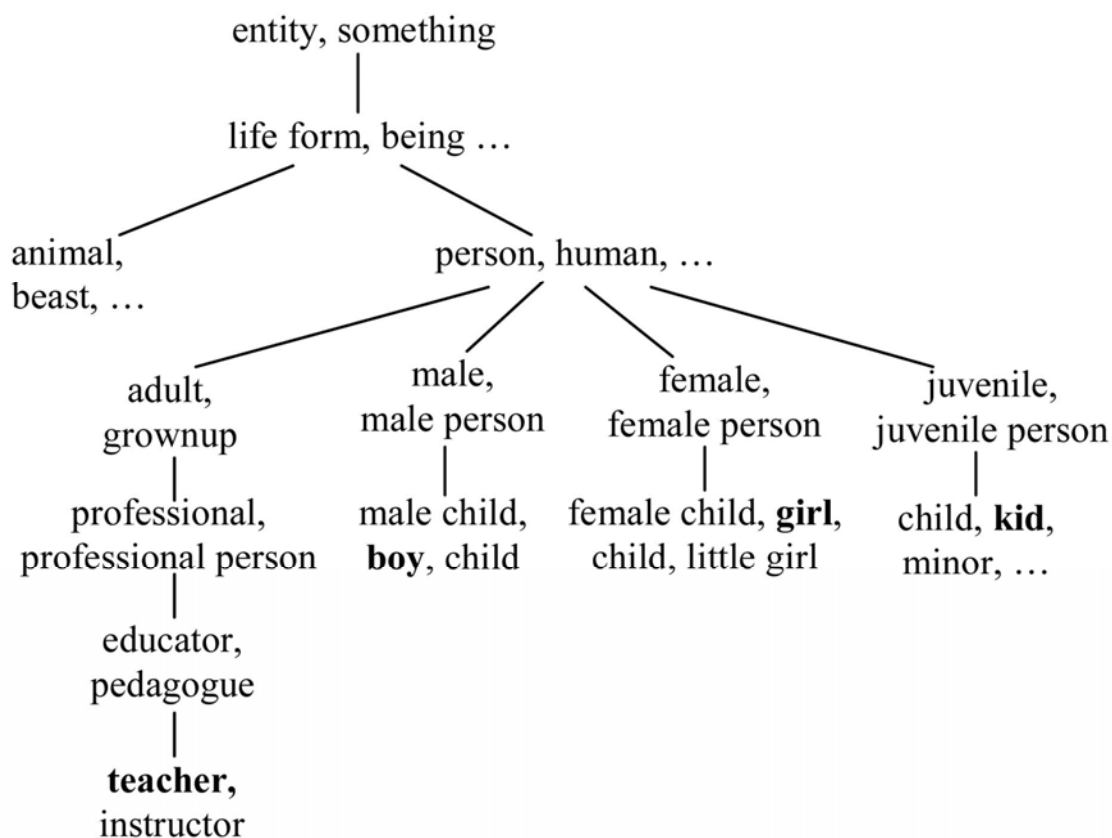
There are two methods to provide an access to a large number of Web documents. The first method is by indexing documents through Web crawling; this has the inherent problems of Web documents that face any Web retrieval system such as bulk size, heterogeneity, and duplication [2], however the system could 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-words 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 hierarchical 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.

## CHAPTER 2

### LITERATURE REVIEW

#### 2.1 Introduction

This chapter firstly reviews some plagiarism detection methods and research prototypes that were covered in the literatures. Those methods came from different areas such as Information Retrieval (IR), Natural Language Processing (NLP), and Data Mining. The discrepancy of this variety of methods is based on the fact that the problem of written text plagiarism can take several forms.

Some terms will be used frequently throughout the rest of this report and are defined here; A *Document*: is a body of text from which structural information can be extracted. A *Corpus*: is a collection of such documents. A *Token*: is any string of alphanumeric text taken from some document, such as a character, word, or sentence. A *Chunk*: is any order of tokens.

## 2.2 Document Plagiarism

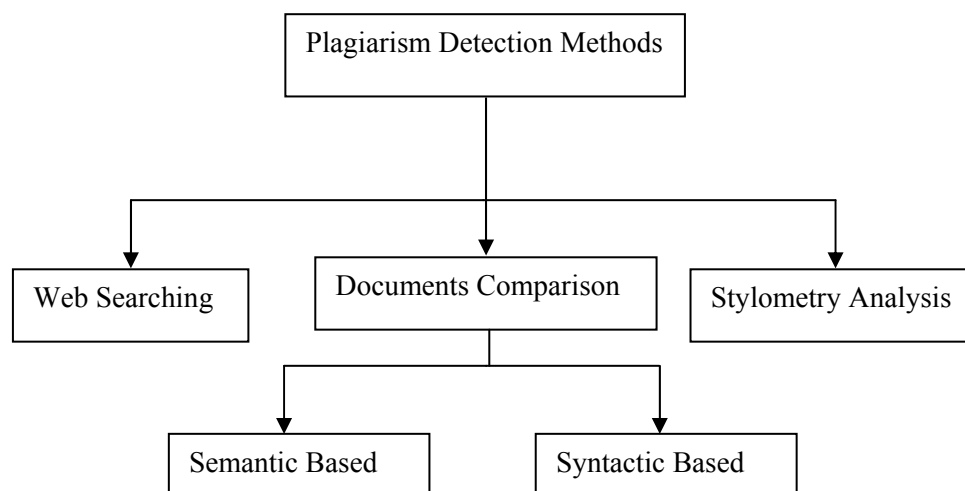
Opposed to other types of plagiarism (such as music, graphs, etc), document plagiarism falls in two categories; source code plagiarism and free text plagiarism. Given the constraints and keywords of programming languages, detecting the former is easier than detecting the latter and hence source code plagiarism detection is not the focus of current research [1].

Plagiarism takes several forms. Maurer et al [3] stated that the followings are some of what considered practices of free text plagiarism:

- Copy-paste: or verbatim (word-for-word) plagiarism, in which the textual contents are copied from one or multiple sources. The copied contents might be modified slightly.
- Paraphrasing: changing grammar, using synonyms of words, re-ordering sentences in original work, or restating same contents in different semantics.
- No proper use of quotation marks: failing to identify exact parts of borrowed contents.
- Misinformation of references: adding references to incorrect or non existing sources.
- Translated Plagiarism: also known as cross-language plagiarism, in which the contents are translated and used without reference to original work.

### 2.3 Plagiarism Detection Methods

Following Maurer et al. [3] plagiarism detection methods can be broadly classified into three main categories; the first category tries to capture the author style of writing and find any inconsistent change in this style. This is known as *Stylometry analysis*. The second category is more commonly used which is based on comparing multiple documents and identifying overlapping parts between these documents. The third category takes a document as input and then searches for plagiarism patterns over the Web either manually or in an automated manner. Figure 2.1 provides taxonomy of plagiarism detection methods.



**Figure 2.1** Taxonomy of plagiarism detection methods

### 2.3.1 Detection based on Stylometry Analysis

In some cases the original documents may not be available. For example, when someone copies some content from a book which is not in a digital format, or when someone else do some work for a student assignment. In this case all plagiarism detection methods that are based on documents comparison are not useful. This problem motivated some researchers to introduce new methods that do not depend on a reference collection.

Detection methods that are applied to one or more documents belong to the same author, and without external sources, are referred as *intrinsic plagiarism detection methods* [3, 13]. The most well-known methods are Stylometry methods. Stylometry is a statistical approach to determine the authorship of literature. This approach requires well defined quantification of linguistic features (known as Stylometric features) which can be used to determine inconsistencies within a document [3]. The intuition behind this class of methods is based on the presumption that every author has a unique style of writing; if this style has changed along with several successive sentences or paragraphs then the document is considered as plagiarized [12]. The plagiarism can be identified, for example, when the author interchangeably use the pronouns “We/our” and “I/my”, or when the style of using prepositions and articles have been changed considerably.

Depending on the chunk size and type, most of Stylometry features fall in one of the following five categories [13]: (i) Text statistics: operate at the character level, (ii) Syntactic features: measure the writing style at the sentence level, (iii) Part-of-speech features: quantify the use of word classes, (iv) Closed-class word sets: count special words, and (v) structural features: which reflect text organization.

The Stylometry approach is not commonly used [3,13] , this is because it is hard to prove plagiarism without evidence from the source documents. Nevertheless this approach could provide an indication to which documents are likely to be plagiarized and therefore used for further comparison.

### **2.3.2 Detection based on Documents Comparison**

The major goal of any plagiarism detection system is to highlight copyright violations. As mentioned in section 2.2, a violation can occur when a fragment of text of whatever size and distribution is duplicated between two or more documents belonging to different authors, in this case the system syntactically searches for any such overlaps. However, due to the complexity of natural languages, it is possible that the same content are presented in different semantics (e.g., paraphrasing), or the same words or phrases could have different meanings in different contexts, in this case a deep analysis must be used by the system, and some Natural Language Processing (NLP) techniques could be employed. In both cases it is required that a referential collection of documents (corpus) exist. This section briefly discusses methods for both semantic and syntactic plagiarism detection.

#### **2.3.2.1 Semantic-Based Detection**

Most copy detection system can only compare syntactically similar words and sentences, thus if the copied materials are modified considerably it is difficult to

detect plagiarism in such systems. The modification can range from replacing words by their synonyms, to introducing the same concept under different semantics.

By using WordNet thesaurus for retrieving synonyms the problem of word substitution could be handled, however because word senses are ambiguous, selection of the correct term is often non-trivial [38].

For more complex plagiarism patterns such as sentence structure changes, a deeper analysis is required [36,37]. Kang et al. introduced the system PPChecker [36] that calculates the amount of data copied from the original document to the query document, based on linguistic plagiarism patterns. Since they used sentences as comparing units between documents, they identified five patterns; the exact sentence copying, word insertion, word deletion, word substitution between sentences, and the whole sentence change pattern. Those patterns are identified based on three decision conditions; word overlap, word difference, and size overlap. For each pattern, they identified different similarity measure and achieved impressive results over some syntactic-based systems. Tachaphetpiboon et al. [37] proposed a novel linguistic analysis method for plagiarism detection, using syntactic-semantic analysis. Syntactic analysis was carried out by the use of a parser to identify grammar rules in the texts and determine the structures of the texts. Then, the structures of the texts are compared by grammar rules. Their system as well as PPChecker used WordNet for retrieving synonyms.

Some methods utilize statistics information such as words' positions in documents to measure their similarity. Bao et al [45] introduced a method called Semantic Sequence Kin (SSK) that considers the word's position information so as to detect plagiarism at fine granularity. They defined semantic sequence in some string  $S$  as a continual word sequence after the low density, where continual means that if two words are adjacent in  $S$ , then the difference between their positions in  $S$  must not be greater than a threshold, and density denote the reciprocal of the difference

between two occurrences of a word in S. their observation was based on that by taking into account the position of each word, then plagiarism can be identified. Later they introduced Common Semantic Sequence Model [46], which is similar to semantic sequence kin model, but uses another formula to calculate similarity of semantic sequences

### 2.3.2.2 Syntactic-Based Detection

Unlike semantic-based, syntactic-based methods do not consider the meaning of words, phrases, or sentence. Thus the two words “exactly” and “equally” are considered different. This is of course a major limitation of these methods in detecting some kinds of plagiarism. Nevertheless they can provide significant speedup gain comparing to semantic-based methods especially for large data sets since the comparison does not involve deeper analysis of the structure and/or the semantics of terms.

To quantify the similarity between chunks, usually a similarity measure is used. As an example, consider the following five chunks where letters represent words.

A B C D E    A F C D E    A B F C D    A B C F D    A B C D F

The underlined words indicate that all five chunks share four words which make them possible instances of plagiarism. Consider now the following similarity function:

$$j(x, y) = \frac{|x \cap y|}{|x \cup y|}$$

Where  $x$  and  $y$  are two sets of words and  $|x|$  is the number of words in  $x$ , every pair of documents in the running example has  $j(x, y) = 4/6$ , indicating that  $x$  and  $y$  share four words out of five.

The previous similarity function is the Jaccard resemblance. Such methods for measuring the similarity between documents were derived from Information Retrieval (IR). Those methods do not give a “yes” or “no” answer to the question of whether the documents are relevant to the user’s need, but orders them by estimated likelihood of relevance [16]. This estimation is captured using a *similarity measure* which normally is a function that takes two subsets of documents as input and produce a value that indicates the similarity between the two documents; documents are then *ranked* according to their similarity value with the query document.

Shivakumar et al [19] introduced the system SCAM and the famous Relative Frequency Model (RFM) which is a modification of the Cosine function. SCAM was demonstrated to perform better than a sentence matching systems named COPS (see section 2.7.2.1 ) in many cases of detecting plagiarism[19], however it produced more false positives (documents that reported as plagiarized, though they are not), in some cases SCAM reported two different documents as being 100% equal. Also since SCAM measures the global similarity it cannot introduce positional information about the copied contents.

Hoad and Zobel [16] considered the problem of identifying coderivative documents; that is documents they originated from the same source. For this purpose they made five variations of the standard Cosine measure in which they call them the Identity Measures. The design of the identity measure was based on the intuition that similar documents should contain similar numbers of occurrences of words. All of the five variations make use of *term weight* which is an expression of the importance of a term in a given document calculated as the frequency of occurrence of that term.

## 2.4 Existing Web-Based Plagiarism Detection Tools

This section reviews some existing plagiarism detection tools and highlights some weaknesses of these tools based on a comparative study on 10 abstracts selected from ACM digital library and manually plagiarized by synonym replacing.

Most Web-based plagiarism detection tools use search engine APIs. An example of such tools is DocCop[48] which is one of the most simple and basic tools. The tool chunks the query document into N-grams (consecutive words of length N) and then uses the grams as queries. It then measures the degree of plagiarism by the percentage of queries with non-empty response from the search engines divided by the number of all queries. Figure 2.2 shows a sample of report generated by DocCop.

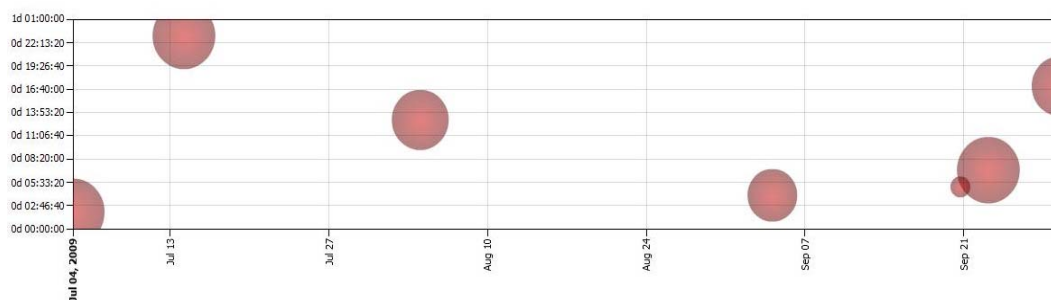
599 of 617	has been with-success incorporated into the ontoln framework, a natural	Google	<input type="checkbox"/>	Yahoo!	<input type="checkbox"/>	MS Bing	<input type="checkbox"/>
600 of 617	been with-success incorporated into the ontoln framework, a natural language	Google	<input type="checkbox"/>	Yahoo!	<input type="checkbox"/>	MS Bing	<input type="checkbox"/>
601 of 617	with-success incorporated into the ontoln framework, a natural language interface	Google	<input type="checkbox"/>	Yahoo!	<input type="checkbox"/>	MS Bing	<input type="checkbox"/>
602 of 617	incorporated into the ontoln framework, a natural language interface generator	Google	<input type="checkbox"/>	Yahoo!	<input type="checkbox"/>	MS Bing	<input type="checkbox"/>
603 of 617	INTO THE ONTONL FRAMEWORK, A NATURAL LANGUAGE INTERFACE GENERATOR FOR	Google	<input type="checkbox"/>	Yahoo!	<input type="checkbox"/>	MS Bing	<input type="checkbox"/>
604 of 617	THE ONTONL FRAMEWORK, A NATURAL LANGUAGE INTERFACE GENERATOR FOR KNOWLEDGE	Google	<input type="checkbox"/>	Yahoo!	<input type="checkbox"/>	MS Bing	<input type="checkbox"/>
605 of 617	ONTONL FRAMEWORK, A NATURAL LANGUAGE INTERFACE GENERATOR FOR KNOWLEDGE REPOSITORIES.	Google	<input type="checkbox"/>	Yahoo!	<input type="checkbox"/>	MS Bing	<input type="checkbox"/>
606 of 617	FRAMEWORK, A NATURAL LANGUAGE INTERFACE GENERATOR FOR KNOWLEDGE REPOSITORIES. THE	Google	<input type="checkbox"/>	Yahoo!	<input type="checkbox"/>	MS Bing	<input type="checkbox"/>
607 of 617	A NATURAL LANGUAGE INTERFACE GENERATOR FOR KNOWLEDGE REPOSITORIES. THE EXPERIMENTATIONS	Google	<input type="checkbox"/>	Yahoo!	<input type="checkbox"/>	MS Bing	<input type="checkbox"/>
608 of 617	natural language interface generator for knowledge repositories. the experimentations demonstrate	Google	<input type="checkbox"/>	Yahoo!	<input type="checkbox"/>	MS Bing	<input type="checkbox"/>
609 of 617	language interface generator for knowledge repositories. the experimentations demonstrate a	Google	<input type="checkbox"/>	Yahoo!	<input type="checkbox"/>	MS Bing	<input type="checkbox"/>

**Figure 2.2** An example HTML report generated by DocCop

When DocCop tested by the 10 plagiarized abstracts it was not able to retrieve any document.

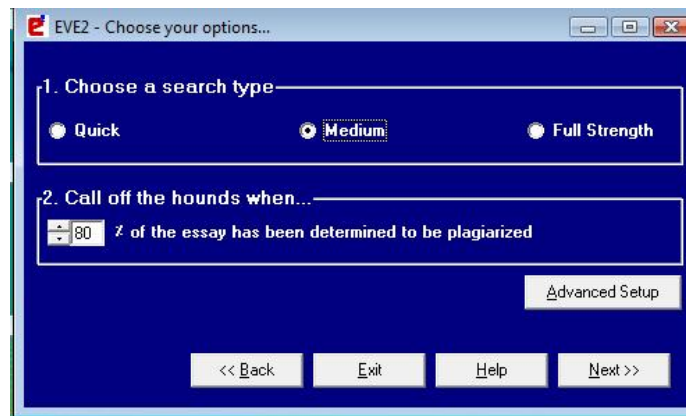
Another freely available tool that is based on search engines API is Plagium[28]. It is not clear the way that Plagium uses, however it performed better

than DocCop in detecting the plagiarized abstracts and was able to retrieve 2 out of the 10 documents. The tool returns a graphical timeline showing the source documents and how much they share information with the query document. Figure 2.3 shows a sample report returned from Plagium.



**Figure 2.3** A sample report with timeline returned from Plagium

Some web-based tools do not depend on search engines APIs. EVE2 [71] is an example of such tools. EVE2 is a commercial tool which allows the user to customize the search as depicted in Figure 2.4. EVE2 claims that it performs extensive searching and the target is any web document. By testing EVE2 with the 10 plagiarized abstracts it always showed a message indicating that it found no instances of plagiarism. It was also tested with full copied document from digital libraries including ACM and IEEE, and also from other sites including Wikipedia but EVE2 failed in retrieving the source documents in all tests.



**Figure 2.4** The interface of EVE2 for Web searching

Turnitin [70] is another commercial tool and perhaps the most famous and successful one[3]. Turnitin uses its own Web index in searching for plagiarism instances. It was not tested in this initial comparative study. Table 2.1 Shows properties of some existing tools based on [59].

**Table 2.1** Properties of some existing plagiarism detection tools based on [59].

	<b>Turnitin</b>	<b>MyDropBox</b>	<b>PAIRwise</b>	<b>EVE2</b>	<b>WCopyFind</b>	<b>CopyCatch</b>
<b>URL</b>	www.turnitin.com	www.mydropbox.com	http://www.pairwise.cits.ucsb.edu/	www.canexus.com	http://plagiarism.phys.virginia.edu/Wsoftware.html	http://www.copycatchgold.com/index.html
<b>Type</b>	Web based	Web based	Web based	Download	Download	Download
<b>Databases</b>	ProQuest	2.7 million articles from ProQuest + 5.5 million from FindArticles	None	None	None	None
<b>Papermills</b>	None	150,000 papers	None	None	None	None
<b>Internet</b>	4.5 billion pages updating 40 million/day	8 billion documents from MSN Search Index	Yes	Only searches the Internet	Only compares submitted papers to each other	CopyCatch Web searches the Web with a Google webapi key.
<b>Submitted papers</b>	10 million previously submitted papers	All previously submitted papers from within same institution	Yes	None	User must provide the documents for comparison against each other.	User must provide the documents for comparison against each other.

## 2.5 Semantic Networks

A semantic network or net “is a graphic notation for representing knowledge in patterns of interconnected nodes and arcs” [50]. The most influential example of such networks in computational linguistics is WordNet[4]. WordNet is a lexical database for English language that organizes words in synonym sets (Synsets) each of which represents a distinct concept. A synset contains synonym words or collocations of words and provide a short textual representation of the synset. An example of a synset is shown in Figure 2.5

**{computer, computing machine, computing device, data processor, electronic computer, information processing system}** (*a machine for performing calculations automatically*)

**Figure 2.5** An example of a synset in WordNet

Synsets are connected by semantic and lexical relations. Table 2.2 shows some of those relations and a brief description about each relation.

**Table 2.2** Some of the relations between concepts in WordNet (N=noun, V=Verb, Adj=adjective, Adv=adverb)

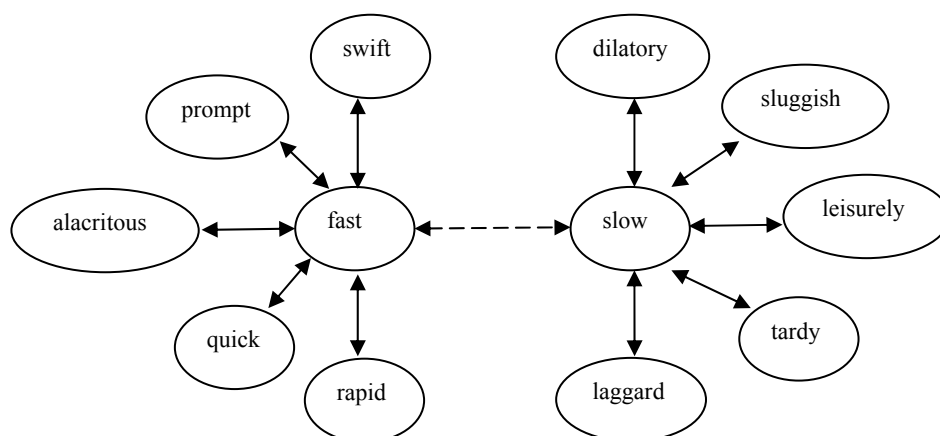
<b>Relation</b>	<b>Description</b>	<b>Applies To</b>
<i>hypernym</i>	<i>Y</i> is a hypernym of <i>X</i> if every <i>X</i> is a (kind of) <i>Y</i>	N-N, V-V
<i>hyponym</i>	<i>Y</i> is a hyponym of <i>X</i> if every <i>Y</i> is a (kind of) <i>X</i>	N-N
<i>coordinate term</i>	<i>Y</i> is a coordinate term of <i>X</i> if <i>X</i> and <i>Y</i> share a hypernym	N-N , V-V
<i>holonym</i>	<i>Y</i> is a holonym of <i>X</i> if <i>X</i> is a part of <i>Y</i>	N-N
<i>meronym:</i>	<i>Y</i> is a meronym of <i>X</i> if <i>Y</i> is a part of <i>X</i>	N-N
<i>troponym:</i>	the verb <i>Y</i> is a troponym of the verb <i>X</i> if the activity <i>Y</i> is doing <i>X</i> in some manner	V-V
<i>entailment:</i>	the verb <i>Y</i> is entailed by <i>X</i> if by doing <i>X</i> you must be doing <i>Y</i>	V-V
<i>Pertainym</i>	e.g.,(biological pertains to biology)	Adj-N
<i>Similar to</i>		Adj-Adj
<i>participle of</i>	e.g.,(elapsed participle of verb elapse)	Adj-V
<i>root adjectives</i>	e.g.,(computational is a root adjective of computationally)	Adv-Adj
<i>Antonym</i>		N-N, V-V, Adj-Adj, Adv-Adv
<i>See also</i>		V-V, Adj-Adj
<i>Attribute</i>		Adj-N

WordNet distinguishes between nouns, verbs, adjectives, and adverbs since they follow different grammatical rules. Table 2.3 shows the number words of each part-of-speech in WordNet 2.1.

**Table 2.3** Statistics about WordNet 2.1

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117,798	82,115	146,312
Verb	11,529	13,767	25,047
Adjective	21,479	18,156	30,002
Adverb	4,481	3,621	5,580
Totals	155,287	117,659	206,941

Nouns and verbs are organized into hierarchies based on the hypernym/hyponym relation between synsets. Adjectives and adverbs, however, do not follow this type of organization. Adjectives are arranged in clusters containing head synsets and satellite synsets. Each cluster is organized around antonymous pairs (and occasionally antonymous triplets). Most head synsets have one or more satellite synsets, each of which represents a concept that is similar in meaning to the concept represented by the head synset. Figure 2.6 shows an example of a bipolar adjective structure.

**Figure 2.6** Bipolar adjective structure (→= similarity, -->=antonymy)

Pertainyms are relational adjectives and do not follow the structure just described. Pertainyms do not have antonyms; the synset for a pertainym most often contains only one word or collocation and a lexical pointer to the noun that the adjective is "pertaining to". Participial adjectives have lexical pointers to the verbs that they are derived from.

WordNet does not have much to say about adverbs. They are not clustered as in the case of adjectives, the organization of adverbs in WordNet is simple and straightforward. Most adverbs are derived from adjectives and have pointers to the adjectives in which they are derived from. Beside this derivation relation, only some adverbs are connected by the antonymy relation.

## **2.6 Document Preprocessing**

A document has to go through several steps before it can be involved in any comparison. Some of these steps are crucial for measuring the overlap between documents. Pre-processing documents is an essential stage before measuring their similarities. Main steps involve tokenization, stop-word removal, and stemming.

### **2.6.1 Tokenization**

The first step in preprocessing is to parse or clean a document by removing irrelevant information, such as punctuation and numbers, remove capitalization and additional spaces. In general a *token* is a unit of a document that may be used by a system. For Web documents it is important to remove document markup such as HTML tags, java script functions, etc. before the documents compared.

### **2.6.2 Stop-word Removal**

Stop-words such as “the”, “of” “and”, etc., indicate the structure of a sentence and the relationships between the concepts presented but do not have any meaning on their own and can be safely removed without effecting the accuracy of measuring how similar two documents is [16,32,33].

### **2.6.3 Stemming**

Many words in the English language have multiple variant forms, distinguished by suffix. The suffixes from variant forms can be removed by stemming [16]. Stemming is not essential step in copy detection but can speed up the process since multiple words are reduced to the same term [16,33,34].

#### 2.6.4 Document Chunking

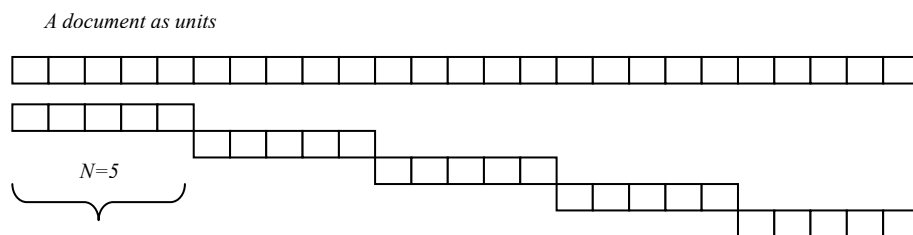
A procedure of breaking a given document into smaller units (tokens) is called *chunking*. The chunking procedure is an important issue in any copy detection system since this procedure will influence the accuracy of the system as well as its performance [19,29].

There are different ways how a document could be chunked[29]:

*Whole document chunking*: the document is trivially a chunk of itself. This method is suitable for detecting near duplicated documents and offers a considerable performance gain, but cannot detect small overlapping as in the case of plagiarized documents.

*Unit chunking*: a document is chunked into smaller units (tokens). The unit could be a character, word, sentence or line. In sentence chunking the document is broken into sentences and then sentences are compared between different documents (e.g., COPS prototype [20]). The main problem here is how to detect the sentence boundary. One approach is to take all words up to a period or a question mark, however sentences that contain abbreviations such as “e.g.” will be broken into multiple sentences due to the embedded periods and the system could fail if there are no discriminative symbols in a given document. Word chunking does not suffer from these limitations since the boundary of words can be identified by whitespace, however the drawback is more false positives since two documents share some words does not mean that a plagiarism has occurred.

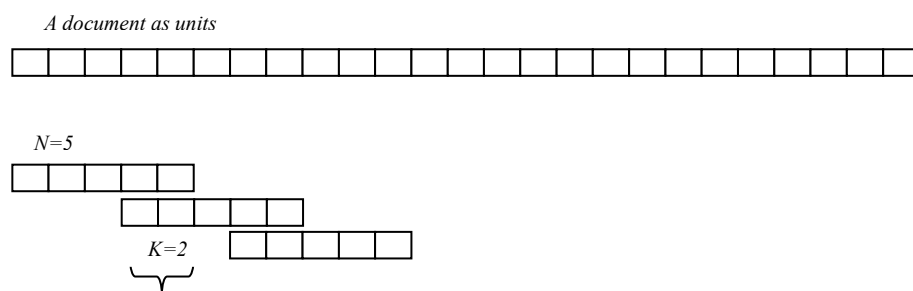
*N-Unit non-overlapped chunking*: in this case the document is broken into *N*-consecutive units (such as characters, words, etc.) using a sliding window with zero overlap between chunks as can be seen from Figure 2.7.



**Figure 2.7** N-Unit non-overlapped chunking strategy with  $N=5$

This method has the advantage of minimizing the candidate that need to be compared as the  $K$  value can be varied depending on the desired comparison level. However a single unit insertion will cause a shifting on the sliding window by one, compromising the accuracy of the detection. When  $N=1$  this method is reduced to a *unit chunking*.

*N-Unit chunking with K-overlap*: here the document is broken into  $k$ -unit chunks, as before, but the chunks overlapped on  $K$  where  $0 < K < N$ . Figure 2.8 depicts this method with  $N=5$  and  $K=2$ .



**Figure 2.8** N-Unit chunking with  $K$ -overlap where  $N=5$  and  $K=2$

*N-grams chunking*: N-gram is a sequence of successive units for a length of N either character-based or word-based. It is a special case of *N-Unit chunking with K-overlap* when  $K=N-1$ . While N-grams character-based chunking is commonly used in typing error detection and Database system integration [39], word-based N-gram chunking is preferred in most plagiarism detection systems due to the ability to capture similar phrases in N-grams since it is difficult to change multiple words for a small length chunking [31,32]. The number of chunks is equal to the number of words in the text, which makes this method the worst in size, but it has the best reliability in finding overlaps [47]

N-grams could be duplicated for some document, removing duplicated N-grams knowing as *shingling*[24]. For example the 4-grams for “A B C A B C A B” are:

{(A,B,C,A); (B,C,A,B); (C,A,B,C); (A,B,C,A); (B,C,A,B)} and the 4-shingles are:  
{(A,B,C,A); (B,C,A,B); (C,A,B,C)}

*Hashed breakpoint chunking*: although the last two chunking strategies reduce the problem of unit shifting, they are not efficient in terms of computation and space cost. another strategy was introduced [20] that reduce the candidates set and takes into account the unit shifting, it works as follows: hash the first unit in the document, if the hash value modulo  $K$  equals zero (for some chosen  $K$ ) then this unit is the first chunk in the document, if not consider the next unit, if its hash value modulo  $K$  equals zero, then the first two units is the first chunk, if not repeat the process until the condition is satisfied and this is a breakpoint. Then the sequence of units from the previous breakpoint until this unit is the chunk.

Clearly the chunking strategy has a tradeoff between the accuracy of measuring overlap between documents and the processing time needed for the comparison. The chunking method should also be the same for all documents[19].

Broder [24] suggested using shingles instead of N-grams for measuring the resemblance and containment of Web documents though the effect of removing duplicated N-grams was not quantified in his work. Liu et al. [40] used a sentence chunking to query a search engine for the application of plagiarism detection; the same chunking was used for the comparison. Tashiro et al [41] used N-unit chunking with K-overlap, for the same purpose, where N, unit, K are 2, 2-words, 2 respectively. They also used the same chunking strategy for both the queried and retrieved documents for the comparison purpose and achieved better precision and recall over sentence chunking. Shivakumar and Garcia-Molina [29] provide a thorough study in comparing different chunking primitives and outlined the relative benefits of these primitives in terms of accuracy and performance over 50,000 documents in which they conclude that the main factor that impacts accuracy is the average length of the chunk. As this length increases it becomes hard to detect partial overlap since overlapping sequences between two documents may start anywhere within the chunk. On the other hand as this length decrease, the loss in chunking sequence may result in false negatives (pairs of documents that identified to have no overlap though they are).

## **2.7 Document Representations and Similarity Measures**

This section details two approaches for representing documents and their corresponding methods for similarity computation. The first approach utilizes semantic networks for deriving features from a document (or parts of documents). The second approach uses document's syntactic information. The two approaches are further detailed in the following two sections.

### 2.7.1 Semantic Based-Representation

The authors in [51] had made an extensive research on methods that used WordNet for deriving the similarity between concepts. They distinguished between three terms; *semantic relatedness*, *semantic distance*, and *similarity*. In their discussion they claimed that similarity is “a special case of semantic relatedness”. An example was given to distinguish between semantic relatedness and similarity is the two words “*cars* and *gasoline*”. The two words are closely more related than “*cars* and *bicycles*”, however the latter pair is more similar. They defined the term semantic distance as the inverse of either semantic similarity or relatedness and stated that “*Two concepts are close to one another if their similarity or their relatedness is high, and otherwise they are distant*”.

In discussing WordNet, the following definitions and notation are used [51]:

- The length of the shortest path in WordNet from synset  $c_i$  to synset  $c_j$  (measured in edges or nodes) is denoted by  $len(c_i, c_j)$ .
- The depth of a node is the length of the path to it from the global root, i.e.,  $depth(c_i) = len(\text{root}, c_i)$ .
- The lowest super-ordinate (or most specific common subsumer) of  $c_1$  and  $c_2$  is denoted by  $lso(c_1, c_2)$ .
- Given any formula  $rel(c_1, c_2)$  for semantic relatedness between two concepts  $c_1$  and  $c_2$ , the relatedness  $rel(w_1, w_2)$  between two words and can be calculated as  $rel(w_1, w_2) = \max_{c_1 \in s(w_1), c_2 \in s(w_2)} [rel(c_1, c_2)]$ . Where  $s(w_i)$  is the set of concepts in the taxonomy that are senses of word. That is, the relatedness of two words is equal to that of the most-related pair of concepts that they denote.

They had compared five approaches of measuring the semantic relatedness between concepts. The first approach [52] makes use of path length and at the same

time considers the weight of the path given by the number of alterations in that path and is given by the following formula for two WordNet concepts  $c1 \neq c2$ :

$$rel(c1, c2) = C - len(c1, c2) - k \times turns(c1, c2)$$

Where  $C$  and  $k$  are constants and  $(c1, c2)$  is the number of times the path between  $c1$  and  $c2$  changes direction.

The second approach [53] is based on the observation “that sibling-concepts deep in a relation appear to be more closely related to one another than those higher up. Each relation has a weight or a range  $[min_r, max_r]$  of weights associated with it. The weight of each edge of type  $r$  from some node  $c_1$  is reduced by a factor that depends on the number of edges,  $edges_r$ , of the same type leaving  $c_1$ ”. This weight is given by the following equation:

$$wt(c1 \rightarrow r) = max_r - \frac{max_r - min_r}{edges_r(c1)}$$

The distance between two adjacent nodes and is then the average of the weights on each direction of the edge, scaled by the depth of the nodes:

$$dist(c1, c2) = \frac{wt(c1 \rightarrow r) + wt(c1 \rightarrow r')}{2 \times \max\{depth(c1), depth(c2)\}}$$

Where  $r$  is the relation that holds between  $c1$  and  $c2$  and  $r'$  is its inverse (i.e., the relation that holds between  $c2$  and  $c1$ ). Finally, the semantic distance between two arbitrary nodes  $c_i$  and  $c_j$  is the sum of the distances between the pairs of adjacent nodes along the shortest path connecting them.

The third approach defines a *conceptual similarity* [54] between a pair of concepts  $c1$  and  $c2$  in a hierarchy by the following equation:

$$\begin{aligned} &sim(c1, c2) \\ &= \frac{2 \times depth(lso(c1, c2))}{len(c1, lso(c1, c2)) + len(c2, lso(c1, c2)) + 2 \times depth(lso(c1, c2))} \end{aligned}$$

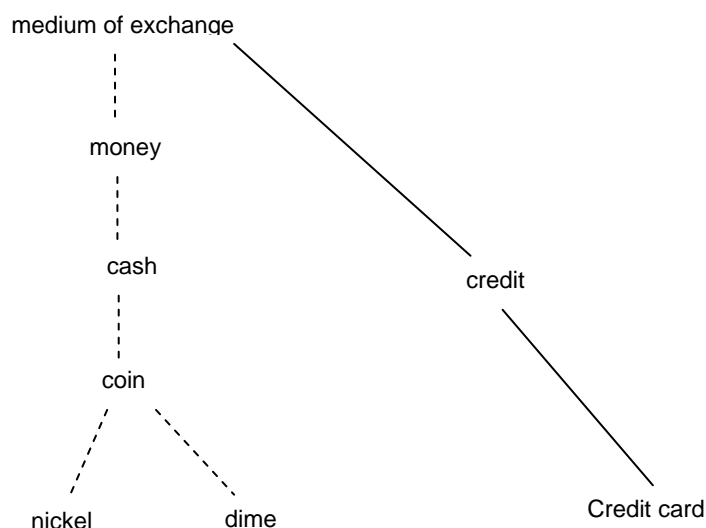
The fourth approach [55] scales the semantic similarity between concepts  $c1$  and  $c2$  in WordNet by the following equation:

$$sim(c1, c2) = -\log \frac{len(c1, c2)}{2 \times \max_{c \in WordNet} depth(c)}$$

The fifth approach is the Resnik's approach [56] in which he call it information content is based on the intuition that one criterion of similarity between two concepts is the extent to which they share information in common. Resnik's information content is defined by the following equation:

$$sim(c1, c2) = \log p(lso(c1, c2))$$

Where  $p(c)$  is the probability of encountering an instance of a concept  $c$ . An example given by Resnik is the difference in the relative positions of the most-specific subsumer of *nickel* and *dime* — *coin* — and that of *nickel* and *credit card* — medium of exchange, as can be seen in Figure 2.9.



**Figure 2.9** An example given by [55] to illustrate the difference between most specific subsumers in WordNet

Note that in the previous discussed methods the similarity was between words and concepts in WordNet. Other methods measure the similarity between sentences. A recently proposed method [57] utilizes most of the previous approaches in deriving the features between word pairs in order to measure the similarity between two sentences. The semantic similarity between two words is given by the following equation:

$$\text{sim}(w_1, w_2) = e^{-\alpha \cdot \text{len}(w_1, w_2)} \cdot \frac{e^{\beta \cdot \text{depth}(\text{lso}(w_1, w_2))} - e^{-\beta \cdot \text{depth}(\text{lso}(w_1, w_2))}}{e^{\beta \cdot \text{depth}(\text{lso}(w_1, w_2))} + e^{-\beta \cdot \text{depth}(\text{lso}(w_1, w_2))}}$$

Where  $\alpha$  and  $\beta$  are constants and used to scale the path and depth respectively. For any two words in the given two sentences the similarity is computed and the maximum similarity is obtained. This maximum similarity is the entry of the semantic vector which is formed from the joint set of word in the sentence pairs. The entry of the semantic vector  $s_i$  is weighted by the following equation:

$$s_i = s'_i \cdot I(w_i) \cdot I(w'_i)$$

Where  $I(w_i)$  and  $I(w'_i)$  are the information contents (Resnik's approach) of a word  $w_i$  in the joint set and its associated word  $w'_i$  in the sentence respectively and is given by the following equation:

$$I(w) = 1 - \frac{\log(n + 1)}{\log(N + 1)}$$

Where  $n$  is the number of occurrence of the word  $w$  in the Brown corpus [58], and  $N$  is the total number of words in that corpus (the corpus contains more than million word).

The overall semantic similarity  $S_s$  between two sentences is measured by the cosine coefficients between their respective semantic vectors:

$$S_s = \frac{s_1 \cdot s_2}{\|s_1\| \cdot \|s_2\|}$$

Where  $s_1$  and  $s_2$  are the semantic vectors of the two sentences.

The algorithm also considers the syntactic similarity between the two sentences. The order similarity  $S_r$  between is obtained by the normalized difference of word order between the two sentences and given by the following equation:

$$S_r = \frac{\|r_1 - r_2\|}{\|r_1 + r_2\|}$$

Where  $r_1$  and  $r_2$  are the order vectors of the two sentences. The order vector is formed in a similar manner to that in the semantic vector except that the entry of the order vector is the relative position of the most similar word in the joint set.

The overall similarity between the two sentences is given by the following equation:

$$S(T1, T2) = \delta.Ss + (1 - \delta)Sr$$

Where  $\delta$  decides the contribution of both semantic ( $Ss$ ) and syntactic ( $Sr$ ) similarities. Based on psychological experiment conducted by [57] the similarity measure performs the best by giving the semantic information a higher weight than the syntactic information, in particular by setting this value to be higher than 80%.

## **2.7.2 Syntactic Based-Representation**

This section introduces three document representations based on syntactic information. Those representations are fingerprinting, term weighting, and N-grams.

### **2.7.2.1 Fingerprinting**

Fingerprinting is the process of creating compact features (fingerprints) of every document in the collection [14,15,16,17]. Two documents are defined to have significant overlap, if they share at least a certain number of fingerprints [15,16,18].

In designing any fingerprinting system for measuring documents similarities there are four issues that need to be considered [16];

*Fingerprint generation:* a fingerprint is generated using a generation function (e.g., MD5 hash function), the function must insure that it will produce the same value for any two equivalent strings and different values for different inputs; this is the core idea of finding similar fingerprints for different documents.

*Fingerprint granularity:* the size of the input to the generation function is known as the granularity of the fingerprint. This granularity must be chosen carefully depending on how two documents to be identified similar or overlapped [17,19,20]. For example if the purpose is to identify near duplicated documents a coarse grained selection can be used, however to identify documents that overlap in sentences or paragraphs, a fine granularity such as sentence granularity, or  $k$ -words granularity for a small  $k$  should be used. Choosing a small granularity such as one word could compromise the accuracy of the detection since two documents are more likely to share some words but not necessary the two documents overlap each other unless some information about the order of words are considered. The choice of granularity also depends on the range of the generation function. For example if the range of the function is 32-bit, then a granularity is chosen such that it will not cause the function to produce hash collisions.

*Fingerprint resolution:* is the number of fingerprints that represent the document. It could be fixed or variable (e.g., based on the document size) depending on the desired storing space and the query evaluation process. Clearly the accuracy of the copy detection depends on the resolution of fingerprints (as well as the other three issues and also depends on the intended application, as mentioned above), for accurate copy detection all generated fingerprints could be used, however, in most practical issues only a subset of the generated fingerprints need to be selected and then stored for the comparison purpose [18].

*Substring selection:* is the strategy of which substrings to be considered. This strategy depends on the fingerprint resolution. If a fixed resolution, say  $n$ , to be produced then  $n$  substring must be selected. There are many alternatives on how the substring to be selected, they can be classified in four classes[16], namely full-fingerprint, positional selection, frequency-based, and structure-based strategies. The full-fingerprint is the simplest and most effective approach [16], in which every substring of length equals to the fingerprint granularity is selected.

The process of fingerprinting a document and subsequently comparing it with other documents' fingerprints is as follows [44]:

1. Partition each document into contiguous chunks of tokens (fingerprint granularity)
2. Retain a relatively small number of representative chunks (fingerprint resolution, substring selection)
3. Digest each retained chunk into a short byte string , each such string called a fingerprint (fingerprint generation)
4. Store the resulting fingerprints in a hash table along with identifying information.
5. If two documents share some fingerprints greater than specified threshold, they are related.

Using Fingerprinting for detecting copyright violations in digital libraries started by Brin et al. [20]. A system called COPS was introduced. COPS used a granularity of one sentence and a variable resolution equals to the number of sentences in the given document. They tested three substring selection strategies, namely the full-fingerprint, overlapped and non overlapped units and hashed breakpoint and found in their experiments that the last one produce good result and save the storing cost.

COPS used a variable resolution, this introduced the problem of favoring large sized documents when compared with the query document. Heintze [17] instead used fixed resolution by selecting phrases producing the lowest hash values in an effort reduce false positives and reduce the storage requirements. Although that approach showed some improvements over variable resolution for a small collection, it was not clear in the experiments if the effect of using other selection strategies and whether this approach can be extended for large datasets.

### 2.7.2.2 Term Weighting Schemes

In Information Retrieval (IR) any document can be represented as vector(s). The content of the vector differ from system to system. The most known representation is the term weighting scheme in the Vector Space Model. Variations of this scheme were also proposed for specific applications. For example [16] made five variations for identifying plagiarized documents. The first variation makes use of difference in term frequencies between two documents and given by the following equation:

$$\frac{1}{1 + |f_d - f_q|} \cdot \sum_{t \in q \cap d} \frac{\log\left(\frac{n}{f_t}\right)}{1 + |f_{d,t} - f_{q,t}|}$$

Where  $f_d$  denotes the number of terms in a document  $d$  and  $f_{d,t}$  is the frequency of the term  $t$  in the document  $d$ . the intuition of using this measure with this weighting scheme is that two plagiarized document, the difference between the frequencies of terms should be small. The second variation is much like the first one but overcome the sensitivity of size difference between two documents by taking the log between the difference and given by the following equation;

$$\frac{1}{1 + \log(1 + |f_d - f_q|)} \cdot \sum_{t \in q \cap d} \frac{\log\left(1 + \frac{n}{f_t}\right)}{1 + |f_{d,t} - f_{q,t}|}$$

The third variation gives a higher rank to documents in which the term is rare in the collection but common in the query or the document by multiplying the term weight by the sum of the frequency of the term in the document and the frequency of the term in the query.

The fourth variation is much like variation two and used to reduce the impact of changing the term weight and given by the following equation

$$\frac{1}{1 + \log(1 + |f_d - f_q|)} \cdot \sum_{t \in q \cap d} \frac{\log\left(\frac{n}{f_t}\right)}{1 + |f_{d,t} - f_{q,t}|}$$

The last variation is same as the previous one but the log operator of the term weight is omitted in order to give rare terms have a much larger weight than common terms. In their experiments they found that the fourth and fifth variations are the best in terms of precision and recall.

### 2.7.2.3 N-Grams

The previous representation makes use of term weight in computing the similarity between documents. Another representation is the N-grams. The value of N can be varied. However with respect to sentences this value should be low. In a recent study dealing with sentence plagiarism detection, Barron, et.al [27] found that

the best values are 2 and 3 (bigrams and trigrams respectively). In this representation a similarity function is used to determine the degree of similarity between sentences. Lane et al. presented the text-plagiarism detector Ferret [21,22,23], where each document is represented as trigrams. Analogously, Bao et al. used Ferret's approach in a study dealing with plagiarism in academic conference papers [25] and Chinese documents [26]. Common similarity measures are shown in Table 2.5, the corresponding binary versions are shown in Table 2.4.

**Table 2.4** Common similarity measures between binary vectors

Cosine function	$sim(x, y) = \frac{\sum x \cdot y}{\sqrt{ x  \cdot  y }}$
Overlap similarity	$sim(x, y) = \frac{\sum x \cdot y}{\min( x ,  y )}$
Dice coefficients	$sim(x, y) = \frac{2 \sum x \cdot y}{ x  +  y }$
Jaccard resemblance	$sim(x, y) = \frac{\sum x \cdot y}{ x  +  y  - \sum x \cdot y}$
Hamming distance	$sim(x, y) =  x  +  y  - 2 \cdot \sum x \cdot y$

**Table 2.5** Common similarity measures between sets

Overlap similarity	$sim(x, y) = \frac{ x \cap y }{\min( x ,  y )}$
Dice coefficients	$sim(x, y) = \frac{2 x \cap y }{ x  +  y }$
Jaccard resemblance	$sim(x, y) = \frac{ x \cap y }{ x \cup y }$
Hamming distance	$sim(x, y) =  (x - y) \cup (y - x) $

## 2.8 Algorithms for Approximate Similarity

An inherent problem in computing most similarity measures is the long time for evaluating the similarity between chunks. This problem was addressed in the database community where it is known as the *similarity join* problem [5, 6, 7, 8, 9, 49] in which the goal is to find all similar pairs of records in large databases. A naïve approach is to compare every pair of records one from each relation which is very inefficient if the size of relations is very large. To solve this problem several algorithms were proposed, those algorithms fall into two categories; signature-based that create a signature for every record such that the intersection between two signature is not empty, followed by a post-filtering process to eliminate false-positives. The second category is based on Information Retrieval, using Inverted Index solutions to minimize the set of considered candidates.

### 2.8.1 Signature Scheme Algorithms

A common framework for signature scheme algorithms is shown in Figure 2.10. The primary difference between these algorithms lies in the scheme used for creating signatures of the input set. The major factor that determines the performance of signature-based algorithms is the number of generated signatures since a large number of signatures means a long processing time required in the post-filtering step.

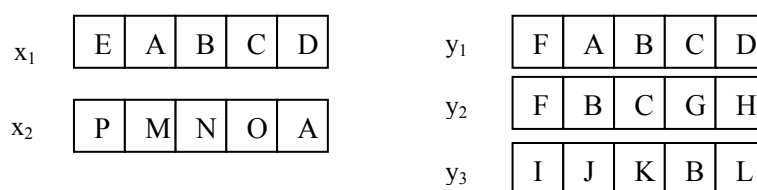
INPUT: two set collections  $R, S$ , similarity function  $Sim(x,y)$  and threshold  $t$

1. For each  $r \in R$ , generate signature-set  $Sign(r)$
2. For each  $s \in S$ , generate signature-set  $Sign(s)$
3. Generate all candidate pair  $(r, s), r \in R, s \in S$ , satisfying  $Sign(r) \cap Sign(s) \neq \emptyset$
4. Output any candidate pair  $(r, s)$  satisfying  $Sim(r, s) \geq t$

**Figure 2.10** Framework for signature-based algorithms [7]

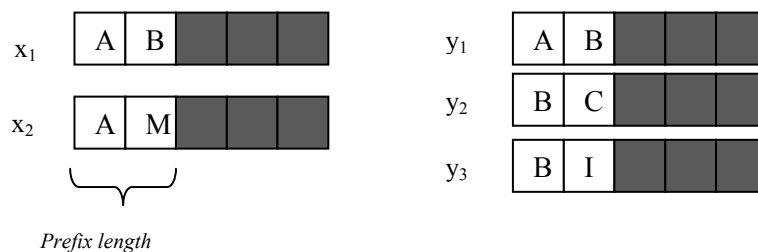
A well-known signature scheme is the *Prefix-Filter* algorithm introduced in [49] for the application of Data Cleaning. It works as follows:

Consider two collections of records  $X = \{x_1, x_2\}$  and  $Y = \{y_1, y_2, y_3\}$  as in Figure 2.11, let the similarity function be the Overlap similarity (Table 2.4), and the similarity threshold be 80%, since all records are of same size  $s=5$ , the Overlap similarity can be written as:  $sim(x, y) = |x \cap y|/5$



**Figure 2.11** Two documents represented as records

For any two records  $x, y$  such that  $x \in X$  and  $y \in Y$ , satisfying  $Overlap(x, y) \geq 80\%$ , the intersection between  $x$  and  $y$  must be greater than 3. Thus instead of measuring the similarity between every pair of records, the records can be sorted and only the first two positions (*prefix-length*) need to be considered as can be seen from Figure 2.12.



**Figure 2.12** Prefix Filter scheme of Figure 2.5 with 80% Overlap similarity threshold

As shown in Figure 2.9,  $x_2$  has no matches and will not be considered, and  $x_1$  has two candidate pairs;  $y_1$  and  $y_2$ . However by consulting the similarity measure,  $y_2$  does not satisfy the condition and consequently ignored in the post-filtering step. The same observation can be extended for other set-similarity measures [49] such as those shown in Table 2.4.

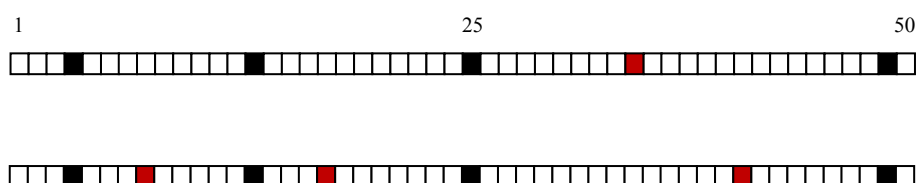
In the previous example the number of comparisons was reduced from 6 to 2 which make the prefix-filtering an efficient scheme to reduce the comparison time. Another efficient algorithm that outperforms Prefix-Filter algorithms is PartEnum [7].

PartEnum is based in the pigeonhole principle and was introduced for Data Cleaning application and using Hamming distance similarity measure. The Hamming distance between two sets is the size of the symmetric difference between the sets. For an illustration consider the following two sets;

$$x = \{A, B, C, D, E\}, y = \{A, B, C, D, F\}$$

The hamming distance  $Hd$  between  $x$  and  $y$  is:  $Hd(x, y) = |(x - y) \cup (y - x)| = 2$ .

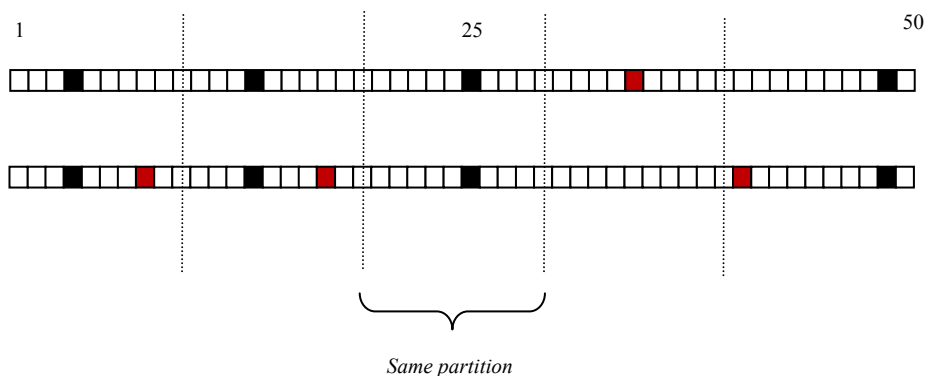
In the same way, the hamming distance between two vectors is the number of dimensions in which the two vectors differ. For example the two binary vectors shown in Figure 2.13 have hamming distance = 4 (shown as red dots) since there are 4 dimensions in which the two vectors differ.



**Figure 2.13** Two vectors with hamming distance = 4

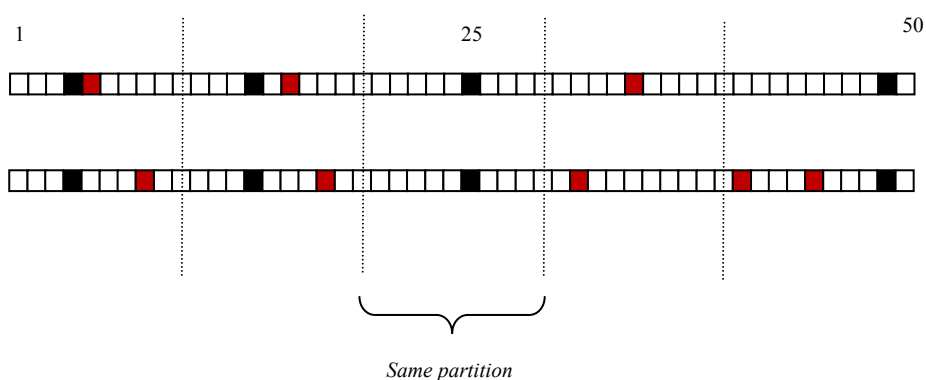
PartEnum is based on two ideas for signature generation; partitioning and enumeration.

*Partitioning*: consider partitioning the domain  $\{1, \dots, n\}$  into  $k + 1$  equi-sized partitions; where  $k$  is the hamming threshold. Any two vectors that have a hamming distance  $\leq k$  must agree on at least one partition, since the number of dimensions in which the two vectors disagree can fall into at most  $k$  partitions. In this case each vector will have  $k+1$  signature. For example let the domain be  $\{1, \dots, 50\}$  and the hamming distance threshold  $k=4$ . As can be seen from Figure 2.14 by partitioning the domain into 5 partitions any two vectors having hamming distance  $\leq 4$  must agree in at least one of these partitions.



**Figure 2.14** Two vectors with hamming distance  $\leq k - 4$  must agree on one of the  $k + 1$  partitions

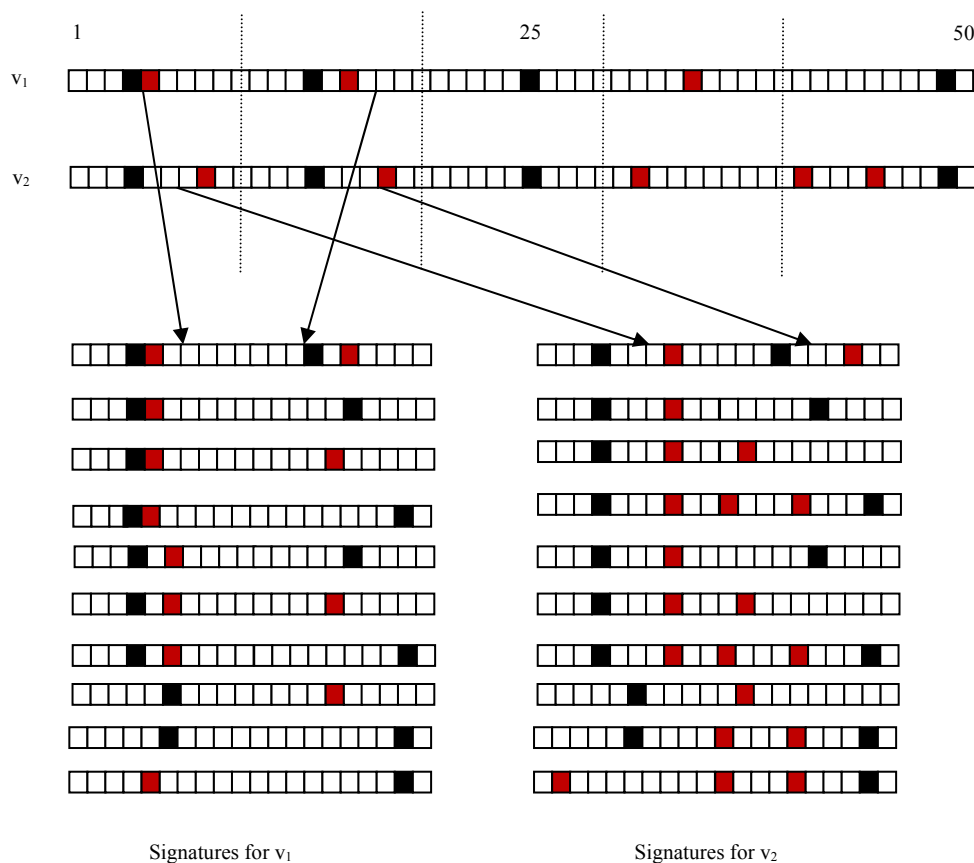
However this scheme is not an effective approach since two vectors often end up accidentally agreeing on one partition even though they are very different which would entail a long post-filtering process to eliminate false-positives. As can be shown from Figure 2.15 the two vectors have one partition in common while the hamming distance between these two vectors is 8.



**Figure 2.15** The two vectors in Figure 2.14 with hamming distance = 8 and agree on one partition

*Enumeration:* In general by partitioning the domain into  $n_2 > k$  equi-sized partitions, where  $k$  is the hamming distance threshold, any two vectors that have a

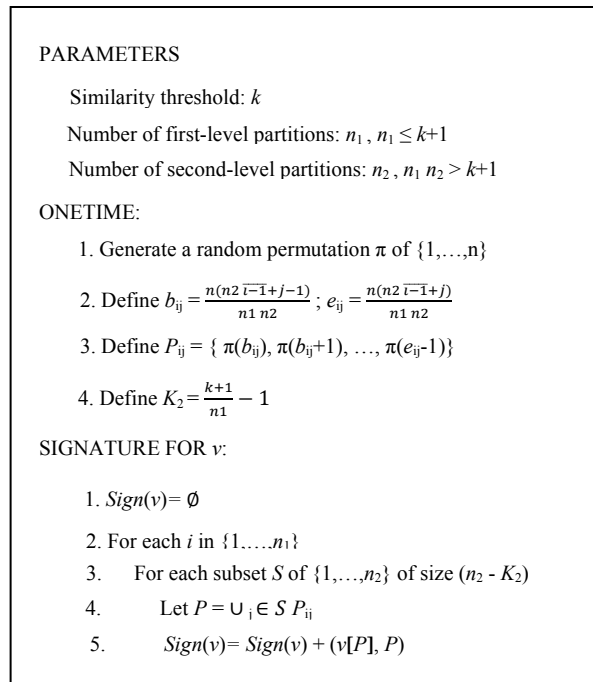
hamming distance  $\leq k$  must agree on at least  $(n_2-k)$  partitions. Using this observation, consider selecting  $(n_2-k)$  partitions in every possible way, any two vectors that have a hamming distance  $\leq k$  must agree on at least one of this selection. Figure 2.16 is an example of this scheme. This example is the same as in Figure 2.15 but with hamming distance threshold=3.



**Figure 2.16** Enumeration scheme for two vectors with hamming distance=3

The two vectors  $v_1$  and  $v_2$  in Figure 2.16 have no signature in common and hence will not be considered. The enumeration scheme has better filtering process but with the drawback of generating large number of signatures for every vector (there are  $\binom{n_2}{k}$  signatures for each vector).

*Hybrid* : PartEnum is a hybrid algorithm that combine both partitioning and enumeration schemes. The formal specification of PartEnum is in Figure 2.17,  $n_1$  and  $n_2$  are two parameters that control the number of signatures generated for each vector. The algorithm first generates a random permutation of the domain  $\{1, \dots, n\}$  and use this permutation to define a two-level partition of the domain (note that from Figure 2.17 the random permutation is generated only once and the signatures for all input vectors are generated using the same permutation). The first-level partition is generated using the partitioning scheme and contains  $n_1$  partitions. And for each first-level partition it generate all possible subsets of size  $(n_2 - k_2)$  using the enumeration scheme. There are  $n_1 \cdot \binom{n_2}{k_2}$  signatures for each input vector. Two vectors are then candidate pair if they have at least one signature in common.



**Figure 2.17** Formal specification of PartEnum[7]

Extension to Jaccard resemblance was covered in [7]. It follows the fact that for two sets  $x$  and  $y$  ,  $k = \frac{(1-t)}{(1+t)} \cdot (|x| + |y|)$  where  $k$  the hamming distance threshold and  $t$  is the Jaccard similarity threshold.

## 2.8.2 Inverted Index-Based Algorithms

Another class of similarity joins algorithms is based on inverted index. The inverted index maps words to the list of record identifiers that contain that word[8]. For vectors the index consists of a number of lists equals to the number of dimensions and each list contains vectors identifiers with non-zero entries in that dimension.

While the main goal of signature-based algorithms lies on minimizing the candidate sets before the post-filtering step, several factors distinguish inverted index-based algorithms and hence determine the performance and scalability of such algorithms [6,8,34,35]. Those primary factors are summarized in Table 2.6.

**Table 2.6** Common factors that influence the performance of Inverted index algorithms

Factor	Description and alternatives
Index Structure	The structure of the index affects directly the scalability of the algorithm since the more parameters of the index, the more main memory consumption.
Index construction	<p>Constructing the index is either:</p> <ul style="list-style-type: none"> <li>- Full construction: scan the data set sequentially and construct a full index for the input sets, the index is then scanned again in one single pass to determine overlapped records.</li> <li>- Dynamic construction: scan the data sets sequentially and overlapped records are determined in the same sequential scan.</li> </ul> <p>The full construction method is not efficient for large datasets since it:</p> <ul style="list-style-type: none"> <li>(i) Fails to utilize some beneficial optimization to minimize the candidate sets such as data sort order and threshold exploitation.</li> <li>(ii) Builds a full index prior to generating any output which results in wasting computation effort.</li> <li>(iii) Requires both the index and input sets remain memory resident for high performance.</li> </ul>
Exploiting the similarity threshold	<p>Exploiting the similarity threshold in an aggressive way can yield dramatic increase in both performance and scalability since not all records satisfy the similarity threshold and hence these records entail unnecessary comparison time and memory storage. Exploiting the threshold can take several forms:</p> <ul style="list-style-type: none"> <li>(ii) During indexing: which means that only those records that have the potential of meeting the similarity threshold need to be indexed.</li> <li>(ii) By using some specific data sort order, such as record size.</li> </ul>

Among others recently proposed algorithms, the All-Pairs algorithm [6] was demonstrated to be highly efficient and outperform previous state-of-art algorithms [5,9,51]. All-Pairs was basically specialized for cosine function and self-join, the formal specification of All-Pairs for binary vectors is shown in Figure 2.18.

```

ALL-PAIRS ( $V, t$ )
1. Reorder the dimensions 1.. $m$  such that the dimensions with the most non-zero entries in  $V$  appear first
2. Sort  $V$  in increasing order of  $|x|$ 
3.  $O \leftarrow \emptyset$ 
4.  $I_1, I_2, \dots, I_m \leftarrow \emptyset$ 
5. for each  $x \in V$  do
6.    $O \leftarrow O \cup \text{Find-Matches}(x, I_1, I_2, \dots, I_m, t)$ 
7.    $b \leftarrow 0$ 
8.   for each  $i$  such that  $x[i] = 1$  in increasing order of  $i$  do
9.      $b \leftarrow b + 1$ 
10.    if  $b/|x| \geq t$  then
11.       $I_i \leftarrow I_i \cup \{x\}$ 
12.       $x[i] \leftarrow 0$ 
13. return  $O$ 

FIND-MATCHES( $x, I_1, I_2, \dots, I_m, t$ )
14.  $A \leftarrow$  empty map from vector id to int
15.  $M \leftarrow \emptyset, \text{remscore} \leftarrow |x|$ 
16.  $\text{minsize} \leftarrow |x| \cdot t^2$ 
17. for each  $i$  such that  $x[i] = 1$  do
18.   Remove all  $y$  from  $I_i$  such that  $|y| < \text{minsize}$ 
19.   for each  $y \in I_i$  do
20.     if  $A[y] \neq 0$  or  $\text{remscore} \geq \text{minsize}$  then
21.        $A[y] \leftarrow A[y] + 1$ 
22.      $\text{remscore} \leftarrow \text{remscore} - 1$ 
23. for each  $y$  with non-zero count in  $A$  do
24.   if  $\frac{A[y] + |y'|}{\sqrt{|x|} \cdot \sqrt{|y|}} \geq t$  then
25.      $d \leftarrow \frac{A[y] + \sum x \cdot y'}{\sqrt{|x|} \cdot \sqrt{|y|}}$ 
26.     if  $d \geq t$  then
27.        $M \leftarrow M \cup \{x, y, d\}$ 
28. return  $M$ 

```

**Figure 2.18** Formal specification of All-Pairs[6]

The algorithm takes a set of vectors  $V$  and a similarity threshold  $t$  and the goal is to find all pairs of vectors  $x, y$  such that  $\cos(x, y) \geq t$ . the top level function scans the dataset and dynamically builds the inverted lists. The FIND-MATCHES function

subroutine scans the inverted lists and perform score accumulation by scanning each list individually.

Besides building the index dynamically and perform the score accumulation, there are three worth noting optimizations were done in this algorithm comparing to a basic inverted index approach. The first optimization is the index reduction (lines 8 through 12). Instead of indexing the whole vector  $y$  the algorithm retains unindexed portion  $y'$  such that  $|y'|/|y| < t$ . correctness for the index reduction follows from the fact that if two vectors  $x$  and  $y$  satisfy the similarity threshold and  $|x| \geq |y|$  (line) then they share at least one term of the indexed portion  $y''$ . This index reduction yields a subtle increase of scalability.

The second optimization (line 18) employs size filtering technique to reduce accesses to inverted lists. It is based on the fact that for any two vectors  $x$  and  $y$  that meet a cosine similarity threshold  $t$ , then  $|y|$  must be greater than or equal  $|x|.t^2$ . Thus any indexed vector  $y$  does not satisfy this *minsize* constraint will not be considered and can be removed or skipped as the algorithm progresses.

The third optimization appears in line 20. The intuition behind this threshold exploitation is that as the algorithm iterates over a vector  $x$ , it get to a point where if a vector has not already been identified as a candidate of  $x$ , then there is no way it can meet the similarity threshold and the algorithm switches to a phase where it avoids putting new candidates in the map (line 21).

## 2.9 Discussion and Summary

Most of the methods [52, 53, 54, 55, 57] that use semantic networks in obtaining the semantic relatedness (or distance) between two concepts  $c_1, c_2$  rely on two semantic attributes:

- The shortest path between two concepts  $len(c_1, c_2)$  measured as the number of edges or nodes in a hierarchical relation that connects the two concepts.
- The depth of the concept that subsumes the two concepts  $len(Iso(c_1, c_2))$  measured as the number of edges or nodes from the uppermost concept in the hierarchy down to the subsumer.

Only Resnik's information content [56] that ignores the two previous attributes and it has been shown in [51] that this approach generated false positives (or suspicious cases) more than those methods that depend on path counting.

An important remark on those methods is that most of the authors had limited their semantic measures to the noun hierarchy in WordNet and in few cases added support to verbs. This is basically due to the fact that adjectives and adverbs are not organized by any hierarchal relation in WordNet. Also most of these methods (except [57]) are word-based rather than sentence-based. However, adjectives and adverbs tend to contribute to the similarity between sentences, and hence should not be ignored.

The algorithms presented in section 2.8 can be viewed as underlying primitives to scale the applicability of N-grams similarities. Table 2.7 highlights the main strength and weakness of both approaches. The All-Pairs algorithm was shown

[6] to consistently outperform PartEnum and other signature schemes by an order of magnitude. This happens since several optimizations can be utilized for using only inverted list approaches.

**Table 2.7** Signature-based versus Inverted index-based algorithms

Algorithm Type	Strength	Weakness
Signature Scheme	<p>(i) The matching between signatures is the identity and can be applied for two documents in a pipeline fashion.</p> <p>(ii) Uses upper and lower bound size-based information to reduce the number of set pairs that need to be considered.</p> <p>(iii) Unlike inverted index approach generating signatures does not require any additional data sorting and consequently additional time.</p>	<p>(i) Generate many signatures for a high dimensionality to ensure completeness.</p> <p>(ii) Fully scans each vector in a candidate pair to compute the similarity score.</p> <p>(iii) Requires parameters tuning (depending on the input size) to scale linearly.</p> <p>(iv) Performance depends on the similarity function as well as the similarity threshold.</p>
Inverted Index	<p>(i) Aggressive in exploiting the similarity threshold and many inputs are never considered.</p> <p>(ii) The index exhibits better memory locality for high dimensionality since only the index is scanned to perform the similarity score and only some dimensions need to be indexed.</p> <p>(iii) Requires no parameters tuning and scale linearly for different input sizes.</p>	<p>(i) Time to build and flush the index is a major disadvantage for multiple comparisons and large datasets.</p> <p>(ii) Performance depends on the similarity function as well as the similarity threshold.</p>

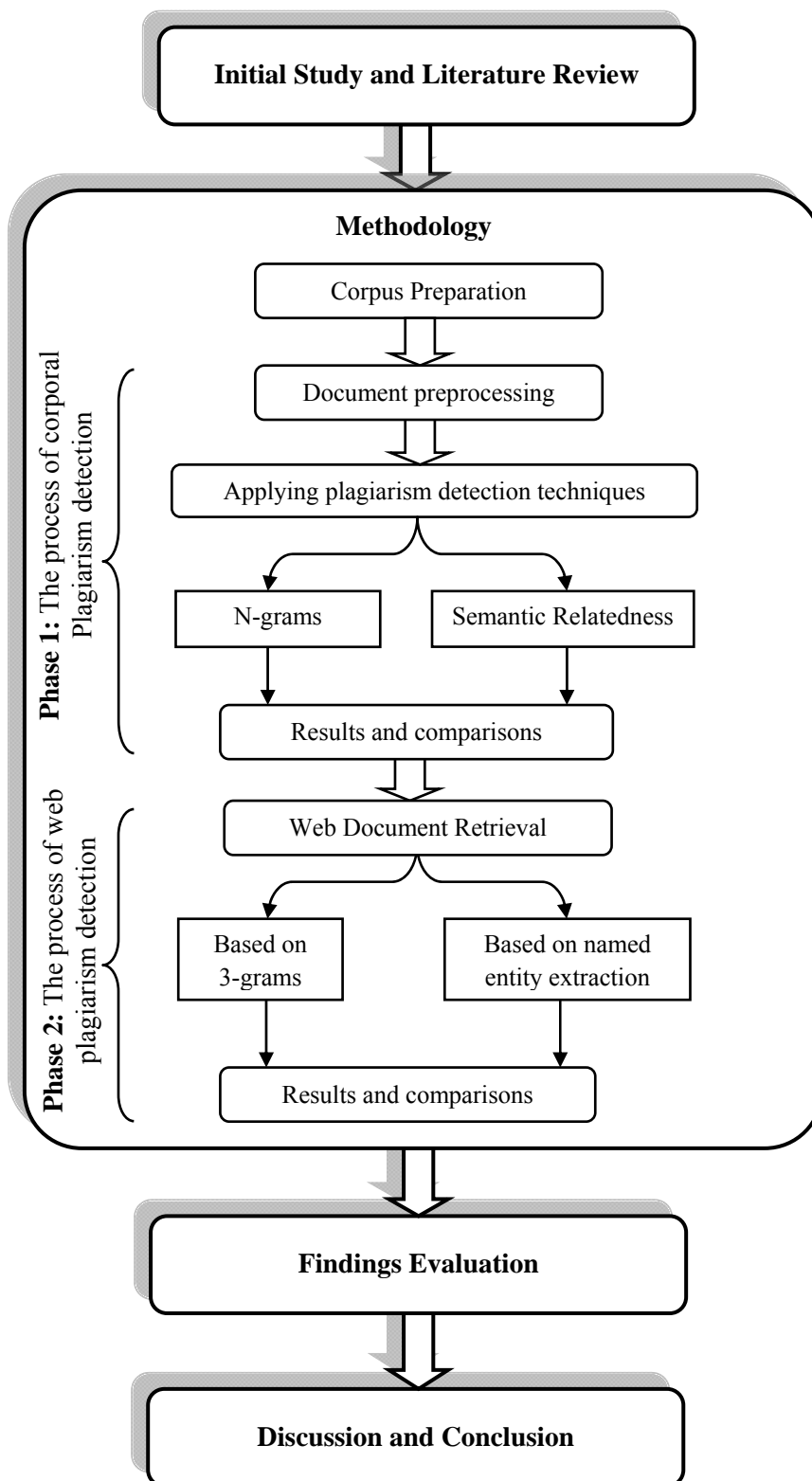
## **CHAPTER 3**

### **METHODOLOGY**

#### **3.1 Introduction**

This chapter presents the methodology of the project. The semantic relatedness approach based on the work of [57] will be adopted in measuring the similarity between sentences by adding supports for other part-of-speeches in particular for adjectives and adverbs. This approach will be evaluated against N-grams representation with three symmetric measures namely Cosine, Jaccard, and Dice coefficients in an inverted index implementation.

### 3.2 Operational Framework



**Figure 3.1** Operational Framework

### **3.2.1 Initial Study and Literature Review**

An initial study was carried during which the basic concepts were gathered. Literatures that related to plagiarism detection, document representation using semantic networks and syntactic information, and similarities measures reviewed in Chapter 2.

### **3.2.2 Corpus Preparation**

Ten documents will be downloaded from ScienceDirect.com [63], those documents constitute the source documents and cover different categories including bioinformatics, software engineering, networking, artificial intelligence and soft computing, and engineering informatics.

From those 10 documents, 20 query documents will be constructed by manually selecting a number of sentences (to be plagiarized) and a record about each sentence will be kept in an information table. Each record in the table has four fields namely, the source document identifier, query document identifier, source sentence identifier, and the query sentence identifier (the identifier of a sentence is its order in a document). This table will be used as a reference in the evaluation phase as will be discussed in section 3.2.7.

Each query document will be differently plagiarized from the others and carries out some or all of the following instances of plagiarism:

- Changing the order of words within a sentence, the structure of sentences.
- Changing words' part-of-speeches (e.g., In computing./the computation..), and inflected forms (e.g., complexity/complexness).
- Removing some of the contents of the original sentence, adding other words in the same context, or adding noisy words.
- Replacing some or most words by their synonyms and antonyms.
- Restating the contents or the idea of a sentence in different meaning, different semantics.
- Only a few numbers of sentences will be left without any change.

Sentences to be plagiarized will be selected manually so that the majority of those sentences can basically cover the aspects of semantic nets and that their words merely support the relations of semantic nets with the focus on the synonymy, similarity, and hypernymy relations. The purpose of this careful selection is twofold. First, we want to assess the contribution of semantic relations in detecting different instances of sentence plagiarism. Second, plagiarists will often select sentences that carry out non-trivial concepts so that it is easy from their point-of-view to hide the original work and hence it is justified to focus on important sentence in a given document rather than sentences with trivial semantics and common senses.

Another 600 documents will be added to the source documents, those documents are from English Wikipedia featured articles [64] that according to Wikipedia “are considered to be the best articles in Wikipedia, as determined by Wikipedia's editors... for their accuracy, neutrality, completeness, and style”. At the time of writing this report there are only 2,612 featured articles, of a total of 3,033,356 articles on the English Wikipedia. The 600 documents are from categories that are similar/different to/from the 10 source documents including computing, engineering and technology, biology, and other categories. The corpus will contains of 610 (original) documents to be compared with 20 query (plagiarized) documents

### 3.2.3 Document Preprocessing

This stage is applied for all query documents as well as the corpus documents. There are four steps in this stage:

- Non-essential tokens such as punctuation, numbers, and parenthesis are excluded. Sentences are extracted during this process by taking all words up to a period or a question mark. Sentences that are less than three words are omitted.
- Stop-words are excluded. The stop-word list is included in Appendix A. Note that the list consists of a small number of words, this is essential since the comparison is between sentences, and sentences are of small length. Those are the words that occur with highest frequency in the Brown corpus. All remaining words are then lowercased. In case of semantic relatedness between sentences, this step is postponed after the part-of-speech tagging step since the tagger will need all information about the processed sentence including the functional words.
- The tokenized, non-stop words are stemmed using the algorithm of [60]. Stemming is applied only to N-grams-based representation. Stemming is not applying when measuring the semantic relatedness between sentences for two reasons. First stemming may reduce words to inflected form so that they might not be found in WordNet. The second reason is to preserve the original meaning of words. Furthermore WordNet has its own morphological analyzer to handle words inflected forms so that they can still be found in WordNet.
- Before measuring the semantic relatedness between sentences, the tokenized words are tagged using the Stanford part-of-speech tagger[61]. The tagger uses the Penn Treebank English POS tag set[62]. There are 36 tags in this set (excluding punctuations) as listed in Appendix D. Some of those tags are mapped to the basic part-of-speech tags that are used in WordNet (noun, verbs, adjectives and adverbs) and the rest of them are discarded. In particular

all functional words such as conjunctions, prepositions, articles, auxiliary verbs, modal verbs, pronouns, and cardinal numbers are removed.

### 3.2.4 Applying Plagiarism Detection Techniques

The procedure of representing documents and the definition of the similarity measure(s) for each technique is detailed in this section.

#### 3.2.4.1 Semantic Relatedness Approach

The algorithm presented in [57] will be adapted in this project to measure the semantic relatedness between two sentence  $T1$ ,  $T2$  as follows:

$$s(w1, w2) = \begin{cases} 1, w1=w2 \\ 0, \text{ no path exists or not of the same POS} \\ e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}}, \text{ otherwise} \end{cases} \quad (3.1)$$

Where  $l$  is the shortest path between  $w_1$  and  $w_2$ ,  $h$  is the subsumer depth,  $\alpha$  scales the effect of path length and equals 0.2, and  $\beta$  scales the depth effect and equals 0.45 [57]. The overall procedure in obtaining the shortest path and the subsumer depth is as follows:

- If the two words are in the same synset the path is set 0. For example, the shortest path between the two nouns *computation* and *calculation* is 0 since they belong to the same synset  $\{computation, calculation, \dots\}$ . The depth in this case is the number of nodes from this synset to the topmost synset
- If the two words are not in the same synset but the two synsets from which they belong contain a common word, the path is set to 1. For example, the path between the two nouns *estimate*  $\{estimate, idea\}$  and *mind*  $\{mind, idea\}$  is set to 1 since the two synsets from which they belong contain a common word *idea*. In this case the depths of the two synsets are calculated and the maximum depth is the relation depth.
- If the two above cases are not presented, the actual path of all word senses is calculated and the shortest one is considered. Once the shortest path is determined the depth of the synset that subsumes the two synsets is the relation depth.

Nouns and verbs in WordNet are organized in hypernym hierarchies but adjective and adverbs are not. In obtaining the shortest path between adjectives or adverbs the first two cases in nouns and verbs (same synset, contain common word) mentioned above are not changed with an exception in setting the depth of the relation. The depth is set to the average depth of the “IS-A” relation in WordNet which equals 6[30]. However in the third case other relations are used in obtaining the path as follows:

- If the two words are adjectives the *similar\_to* relation is consulted to check whether the two synsets from which the two words belong are in the same cluster, if so the path is set to 1 and the average depth is used. If the two synsets are not connected by this relation, both the *pertain\_to* and the *participle\_of* relations are consulted to check whether the two adjectives are pertains to nouns or participle to verbs. If so the same procedure in computing the path and depth between nouns and verbs is applied.
- In case of adverbs the relation *root\_adjectives* is consulted to traverse from adverbs to their root adjectives (if they do have such roots) and the same procedure of adjectives is applied.

As an example in case of adjectives, consider the two adjectives *chemical* and *molecular*. *Chemical* has two senses {*chemical, chemic*} and {*chemical*}, *molecular* has two senses both of them are {*molecular*}. In all senses the two adjectives are not synonymous (within a same synset), do not contain a common word nor they are connected by the *similar\_to* relation, hence the *pertain\_to* relation is used. The adjective *chemical* pertains to the noun chemistry {*chemistry, chemical science*} and the adjective *molecular* pertains to the noun molecule {*molecule*}. The followings are the hypernym trees of the two nouns in WordNet 2.1:

```

{chemistry, chemical science}
=>{natural science}
=>{science, scientific discipline}
=>{discipline, subject, subject area,..}
=>{knowledge domain, knowledge base}
=>{content, cognitive content,..}
=>{cognition, knowledge, noesis}
=>{psychological feature}
=>{abstraction}
=>{abstract entity}
=>{entity}.

```

$\{molecule\}$   
 $\Rightarrow \{unit, building\ block\}$   
 $\Rightarrow \{thing\}$   
 $\Rightarrow \{physical\ entity\}$   
 $\Rightarrow \{entity\}$ .

Hence the shortest path between the two noun senses is 14, the shortest path between the two adjectives *chemical*, *molecular* is 16 (14+2) and the subsumer depth is 1.

An example of adverbs is *significantly* (3 senses) and *considerably* (1 sense), again they are not within a same synset nor they contain a common word. *Significant*, *considerable* are the root adjectives of *significantly*, *considerably* respectively. The two adjectives are connected through the *similar\_to* relation  $\{significant, substantial\} \rightarrow \{considerable\}$ , hence the shortest path between the two adverbs is 1 and the depth equals 6. Figure 3.2 partially illustrates the above mentioned cases in obtaining the semantic attributes.

After applying equation 3.1 to all word pairs in  $T1$  and  $T2$ , the semantic vectors  $s1, s2$  and order vectors  $r1, r2$  are obtained from the joint set of words in both  $T1$  and  $T2$ . The vectors length equals to the size of the joint set. An entry of the semantic vector is given by the following equation:

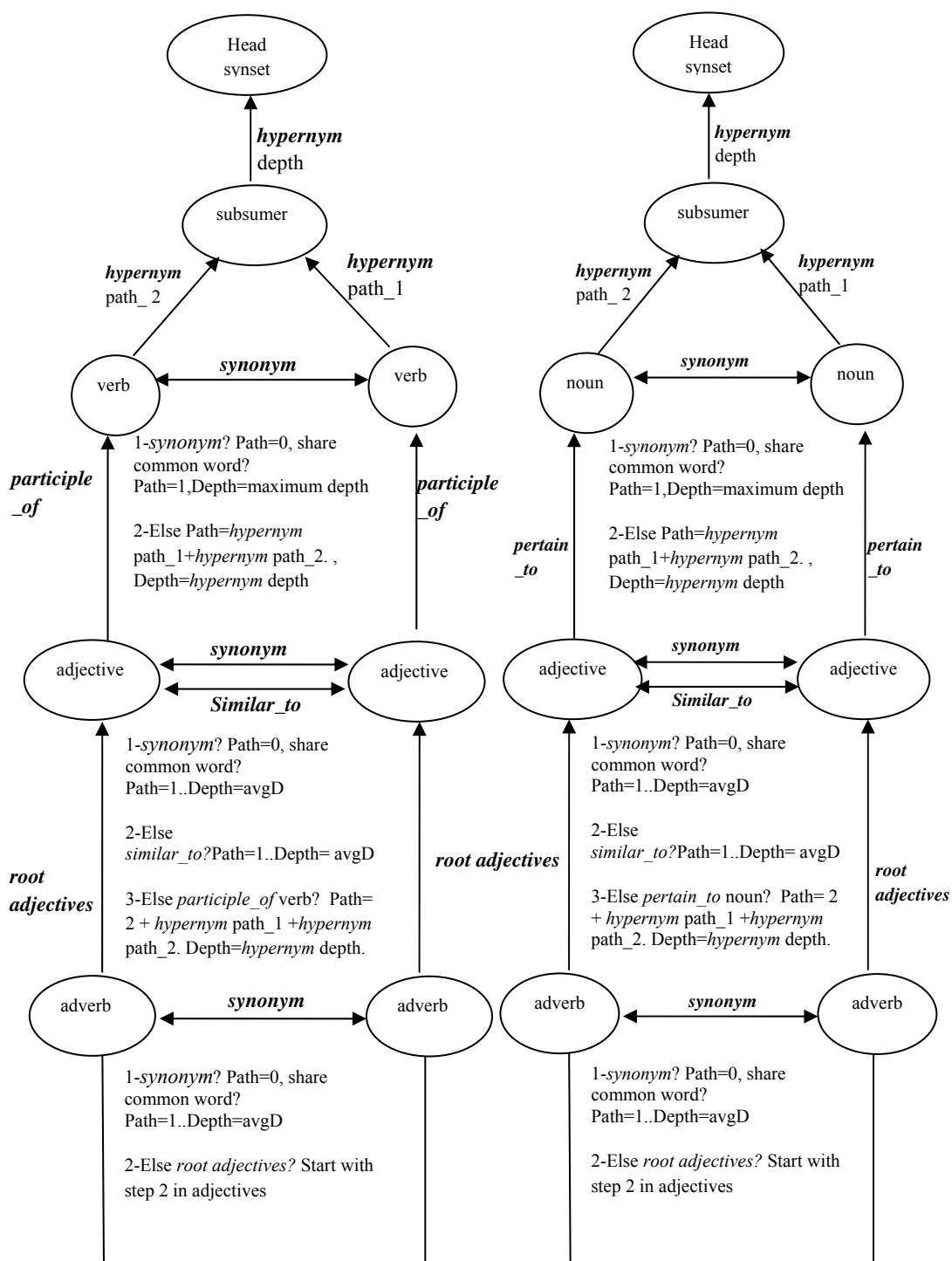
$$s[i] = s'[i].I(wi).I(w'i) \quad (3.2)$$

Where  $wi$  is a word in the joint set and  $w'i$  is its associated word in the sentence obtained as the maximum similar word to  $wi$  based on equation 3.1,  $s'[i] = equation\ 3.1(wi, w'i)$ , and  $I(w)$  is the information content of  $w$  derived from the

Brown corpus [58] as the probability of occurrence of that word in the Brown corpus and given by the following equation:

$$I(w) = 1 - \frac{\log(n + 1)}{\log(N + 1)} \quad (3.3)$$

Where  $n$  is the number of occurrence of  $w$  in the corpus and  $N$  is the total number of words in the corpus. The values of the entries in the semantic vector must exceed the *semantic threshold* which is set to 0.2 [57].



**Figure 3.2** The procedure used in obtaining the semantic attributes between two concepts

The semantic similarity between two sentences is given by the cosine coefficients between their semantic vectors:

$$S_s = \frac{s_1 \cdot s_2}{\|s_1\| \cdot \|s_2\|} \quad (3.4)$$

The order similarity between two sentences is given by the normalized difference between their order vectors (equation 3.5). An entry of the order vector is set to the relative position of the maximum similar word in the sentence to that in the joint set. This entry value must exceed the *order threshold* in order to be considered in the order vector. This threshold is to be optimized in the project as will be discussed shortly.

$$S_r = 1 - \frac{\|r_1 - r_2\|}{\|r_1 + r_2\|} \quad (3.5)$$

Finally the overall similarity between two sentences is given by the following equation where  $\delta$  decides the contribution of semantic similarity and order similarity

$$S(T_1, T_2) = \delta \cdot S_s + (1 - \delta) S_r \quad (3.6)$$

An important consideration is the parameter settings in this representation. The values of  $\alpha$ ,  $\beta$ , and semantic threshold have been optimized for WordNet in [57]. The value of  $\delta$  and the order threshold are responsible for deciding the effect of syntactic information to the similarity between a sentence pair. It is agreed that a common practice of plagiarists is changing the order of words and structure of sentences and hence the two parameters will be optimized in this project for the

application of plagiarism detection. Figure 3.3 shows the pseudo-code of the algorithm.

```

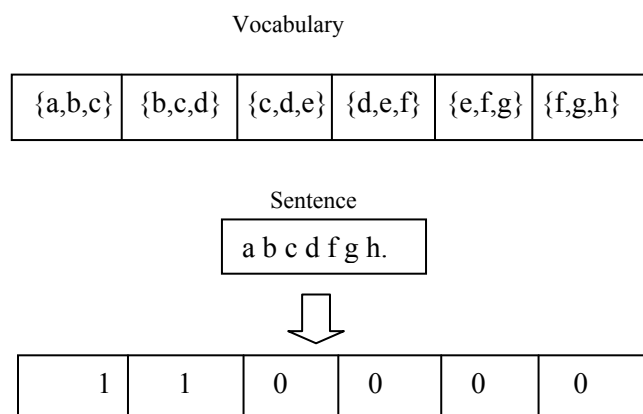
The algorithm of semantic relatedness between a pair of sentences
1. INPUT:
2.   –  $Q$ , is a preprocessed and tagged query sentence.
3.   –  $C$ , is a preprocessed and tagged corpus sentence.
4. PARAMETERS
5.   –  $ts$ , is the semantic threshold.  $tr$ , is the order threshold.  $\delta$ , decides the contribution of semantic
6.     and order information between  $Q$  and  $C$ .
7. OUTPUT
8.   –  $S$  is the semantic relatedness between  $Q$  and  $C$ .
9. BEGIN
10.   $J \leftarrow$  Joint Set of words in both  $Q$  and  $C$ .
11.   $W1, W2 \leftarrow \emptyset$ . // Empty lists to store the associated words used to compute Information Contents.
12.   $r1, r2 \leftarrow$  Empty order vectors that represent  $Q$  and  $C$  respectively of length= $|J|$ 
13.   $s'1, s'2 \leftarrow$  Empty raw semantic vectors that represent  $Q$  and  $C$  respectively of length= $|J|$ .
14.   $s1, s2 \leftarrow$  Empty semantic vectors that represent  $Q$  and  $C$  respectively of length= $|J|$ .
15.  for each word  $j_i \in J$  do //obtaining the raw semantic vectors and order vectors
16.     $s \leftarrow 0, r \leftarrow 0$ 
17.    for each word  $q_k \in Q$  do
18.       $t \leftarrow$  equation 3.1( $q_k, j_i$ )
19.      if  $t > ts$  then
20.        if  $t \geq ts$  then
21.           $s \leftarrow t$ 
22.           $W1[i] \leftarrow q_k$ 
23.          if  $t \geq tr$  then
24.             $r \leftarrow i$ 
25.         $r1[i] \leftarrow r, s'1[i] \leftarrow s$ 
26.       $s \leftarrow 0, r \leftarrow 0$ 
27.      for each word  $c_k \in C$  do
28.        do the same process form line 16 to 23 to obtain  $W2, r2, s'2$ 
29.    for  $i=1$  to  $|J|$  do
30.       $sI[i] = s'1[i].I(W1[i]).I(J[i]) \dots s'2[i] = s'2[i].I(W2[i]).I(J[i])$  //  $I(w)$  is equation 3.3
31.     $Ss \leftarrow$  equation 3.4 ( $s1, s2$ )
32.     $Sr \leftarrow$  equation 3.5 ( $r1, r2$ )
33.     $S \leftarrow \delta.Ss + (1 - \delta)Sr$  //equation 3.6
34.  output  $S$ .
35. END

```

**Figure 3.3** The algorithm for semantic relatedness between a pair of sentences

### 3.2.4.2 N-grams Approach

This representation is defined as follows. The set of N-grams is obtained from the pre-processed query document, each N-gram then correspond to one dimension in the space. Given the set of N-grams  $G=\{g_1,g_2,\dots,g_m\}$ , each sentence  $s$  that either belongs to the query document or a corpus document is an  $m$ -dimensional binary vector  $v$  such that  $v[i]=1$  if  $g_i \in s$ , and  $v[i]=0$  otherwise. Figure 3.4 gives an example of converting a sentence to a binary vector based on 3-grams representation.



**Figure 3.4** Binary vector representation of a sentence

There are three similarity measures that will be evaluated; Cosine, Jaccard, and Dice coefficients. The All-Pairs algorithm [6] was chosen to speed up the process of comparing documents. Figure 3.5 shows the Pseudo-code for cosine similarity as it was introduced in [6].

```

All-Pairs-Cosine
1. INPUT:
2.   -  $R$ , is a collection of binary vectors of length  $n$  represents the query document as inverted
   lists  $I_1, I_2, \dots, I_n$ . Each  $I_i$  maps to all vectors  $r \in R$  such that  $r[i]=1$ .
3.   -  $S$ , is a collection of binary vectors of length  $n$  represents a Web document
4.   -  $t$ , is the similarity threshold.
5. OUTPUT
6.   - All pairs of vectors  $O(r,s)$  satisfying the similarity threshold,  $r \in R$  and  $s \in S$ 
7. for each  $s \in S$  do
8.   |  $O \leftarrow O \cup \text{Find-Matches-Cosine}(s, I_1, I_2, \dots, I_n, t)$ 
9. return  $O$ 

Find-Matches-Cosine( $s, I_1, I_2, \dots, I_n, t$ )
10.  $A \leftarrow$  empty map from vector id to int
11.  $M \leftarrow \emptyset$ ,  $remscore \leftarrow |s|$ 
12.  $minsize \leftarrow |s| \cdot t^2$ 
13. for each  $i$  such that  $s[i] = 1$  do
14.   | for each  $r \in I_i$  such that  $|r| \geq minsize$  do
15.     | if  $A[r] \neq 0$  or  $remscore \geq minsize$  then
16.       |  $A[r] \leftarrow A[r] + 1$ 
17.     |  $remscore \leftarrow remscore - 1$ 
18. for each  $r$  with non-zero count in  $A$  do
19.   |  $d \leftarrow \frac{A[r]}{\sqrt{|r|} \cdot \sqrt{|s|}}$ 
20.   | if  $d \geq t$  then
21.     |  $M \leftarrow M \cup \{r, s, d\}$ 
22. return  $M$ 

```

**Figure 3.5** An inverted index implementation for Cosine similarity[6]

Figure 3.5 is a modification of the original algorithm (Figure 2.18). The dynamic building of the inverted index in All-Pairs (lines 8 through 12 of Figure 2.18) is omitted since the comparison is between two collections of vectors, so either the query document's vectors or the Web documents' vectors will be indexed.

Indexing the former is the choice since (i) it wastes the computation time to build and flush the index every time some Web document is compared with query

document. (ii) The query document's vectors need not to be memory resident in this case since the similarity score can be computed directly from the inverted index. In fact the only needed attribute to compute the similarity score is the size of each vector in the query document.

The minsize (lower-bound) constraint (line 12 of Figure 3.5) is for cosine similarity and remains unchanged in the case of cosine. It follows from the fact that for any pair of binary vectors  $r$  and  $s$  that meets the cosine similarity threshold  $t$ , the following condition must hold:

$$|r| \geq |s| \cdot t^2 \dots [6] \tag{3.7}$$

Where  $t$  is the cosine similarity threshold, and  $|x|$  is the size of vector  $x$  which denotes the number of non-zero dimensions. Extensions to other similarity measures are as follows;

For Jaccard similarity, the following is the minsize constraint between any two vectors  $r$  and  $s$ :

$$|r| \geq |s| \cdot t \tag{3.8}$$

where  $t$  is the Jaccard similarity threshold.

Correctness for this condition is as follows: by definition the Jaccard similarity between two binary vectors  $r$  and  $s$  is :  $Jaccard(r, s) = \frac{\sum r.s}{|r|+|s|-\sum r.s} \geq t$ . Since  $\sum r.s \leq |r|$  we must have that  $\frac{|r|}{|r|+|s|-\sum r.s} \geq t$ , and finally  $|r| \geq |s|.t$ .

Analogously for Dice coefficients the following minsize constraint will be used:

$$|r| \geq \frac{|s|.t}{2} \quad (3.9)$$

Where  $t$  is the Dice threshold. So the two changes in Figure 3.5 are in the Find-Matches-Cosine procedure, in particular line 12 where the minsize constraint for the cosine similarity is replaced by the corresponding minsize constraint for each similarity measure based on equations 3.8 and 3.9, and in line 19 when computing the similarity measure. Figure 3.6 and Figure 3.7 show the Find-Matches procedures for Jaccard, and Dice respectively.

```

Find-Matches-Jaccard( $s, I_1, I_2, \dots, I_n, t$ )
10.  $A \leftarrow$  empty map from vector id to int
11.  $M \leftarrow \emptyset$ ,  $remscore \leftarrow |s|$ 
12.  $minsize \leftarrow |s| \cdot t$ 
13. for each  $i$  such that  $s[i] = 1$  do
14.     for each  $r \in I_i$  such that  $|r| \geq minsize$  do
15.         if  $A[r] \neq 0$  or  $remscore \geq minsize$  then
16.              $A[r] \leftarrow A[r] + 1$ 
17.          $remscore \leftarrow remscore - 1$ 
18. for each  $r$  with non-zero count in  $A$  do
19.      $d \leftarrow \frac{A[r]}{|r| + |s| - A[r]}$ 
20.     if  $d \geq t$  then
21.          $M \leftarrow M \cup \{r, s, d\}$ 
22. return  $M$ 

```

**Figure 3.6** An inverted index implementation for Jaccard Similarity

```

Find-Matches-Dice( $s, I_1, I_2, \dots, I_n, t$ )
10.  $A \leftarrow$  empty map from vector id to int
11.  $M \leftarrow \emptyset$ ,  $remscore \leftarrow |s|$ 
12.  $minsize \leftarrow |s| \cdot t/2$ 
13. for each  $i$  such that  $s[i] = 1$  do
14.     for each  $r \in I_i$  such that  $|r| \geq minsize$  do
15.         if  $A[r] \neq 0$  or  $remscore \geq minsize$  then
16.              $A[r] \leftarrow A[r] + 1$ 
17.          $remscore \leftarrow remscore - 1$ 
18. for each  $r$  with non-zero count in  $A$  do
19.      $d \leftarrow \frac{2 \cdot A[r]}{|r| + |s|}$ 
20.     if  $d \geq t$  then
21.          $M \leftarrow M \cup \{r, s, d\}$ 
22. return  $M$ 

```

**Figure 3.7** An inverted index implementation for Dice coefficients

### 3.2.5 Web Document Retrieval

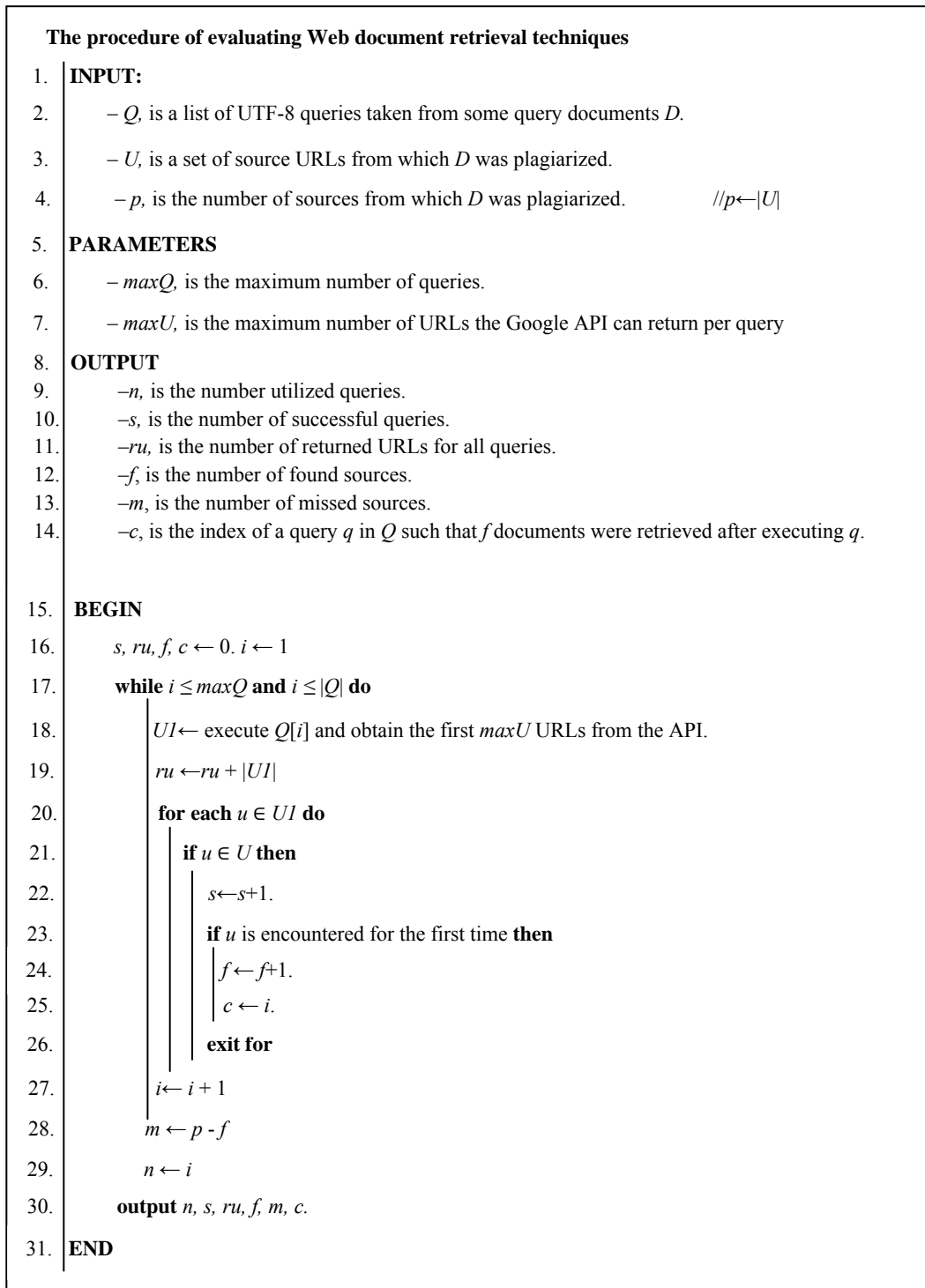
After applying the offline comparison stage, the next stage is the procedure of retrieving the source document from the Web. Each source document's URL will be recorded and the objective is to determine the best technique for retrieving this URL based on the following metrics:

- The number of successful queries over the total number of queries
- The minimum number of queries required to retrieve the source document.

- The number of URLs returned from all queries.
- The number of source documents successfully retrieved.
- The number of missed source documents.
- The number of overall utilized queries by a technique.

There are three techniques that will be evaluated. The first technique takes every n-consecutive words (for some n greater than 2) from the source document as queries. Queries that are totally stop words will be excluded. The value of N is set to 3. The second technique is much similar to the previous one but with a major difference in that the queries are ranked according their importance (weights). Each word in the query is weighted according to equation 3.3. The query weight is the summation of all its individual words' weights.

The third technique is based on extracting named entities and proper nouns since those are usually hard to plagiarize. The extracted entities and nouns are then formulated in sub queries in a decreasing length with a minimum length of two. Figure 3.8 show the procedure for evaluating the three techniques.



**Figure 3.8** The procedure of evaluating Web document retrieval techniques

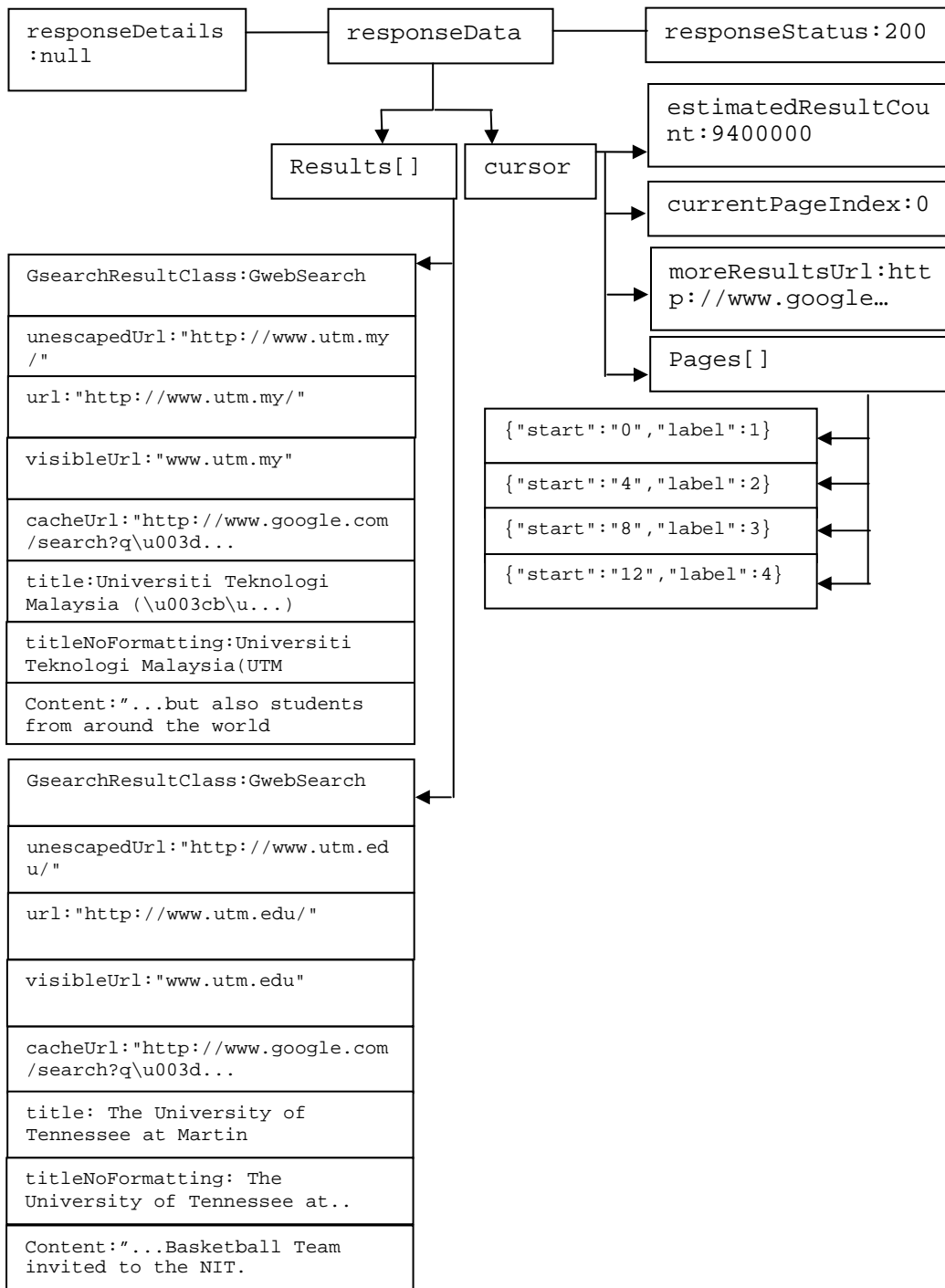
### 3.2.6 Implementation

The implementation phase will be carried out during Project 2 with a 2.0 GHZ Intel Dual Core PC , 4.0 Gigabyte RAM and a 160 Gigabyte-5400 RPM-SATA hard drive. Table 3.1 lists the main libraries that will be used in the experiments and their roles, all libraries are java-based.

**Table 3.1:** Integrated libraries in the project and their roles

Library Name	Its use
Stanford POS (Part-Of-Speech) Tagger [61]	Tagging documents and identifying part-of-speech classes.
JWNL (Java WordNet Library) [69]	Performing the morphological analyzes, accessing WordNet.
Stanford NER (Named Entity Recognizer)[67]	Extracting named entities from query documents.
Google AJAX Web Search API [65]	Web document retrieval.

The Stanford POS, Stanford NER, and JWNL are all open source libraries. The versions that will be used in this project are 1.6, 1.1, and 1.4 respectively which they are the latest versions at the time of writing this report. The Google AJAX API comes with different formats depending on the programming environment. The one that will be used in this project is the *Flash and other Non-Javascript Environments* API. The API exposes a RESTful interface, the method supported is GET and the response format is a JSON object [43] which is very similar to the results obtained from the main Google portal[11]. There is no restriction in the API documentation on the number of queries for a particular period of time. Google, however, limits the results to 64 per query. Figure 3.9 shows the response format of querying the API for: utm.my: <http://ajax.googleapis.com/ajax/services/search/web?v=1.0&q=utm.my>



**Figure 3.9** Response format from querying the Google API

### 3.2.7 Findings Evaluation

Every sentence in a query document will be compared with every sentence in the corpus and the maximum similar one to that query sentence is returned together with the corresponding similarity score. Information about the sentence pair and the similarity score are recorded and compared to the information table that was created during the corpus preparation stage. The followings are the standard metrics in information retrieval that will be used in the evaluation:

$$\text{Recall rate} = \frac{\text{number of true positives}}{\text{number of plagiarized sentences}}$$

If the retrieved sentence is the original sentence then the retrieved sentence is considered as *true positive* otherwise the query sentence is a *false negative*.

$$\text{Precision} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false positives}}$$

When Precision is used and the retrieved sentence is not the original sentence, a manual check is performed between the original sentence and the retrieved sentence. If the query sentence is more similar to the retrieved sentence than to the original sentence, the retrieved sentence is considered as true positive, otherwise the query sentence is a false negative and the retrieved sentence is a *false positive*. In standard Information Retrieval, Precision is accompanied with Recall. Recall is defined as follows:

$$Recall = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

And finally the harmonic mean or F-measure, which gives a single numeric representation of both Precision and Recall, and defined as follows:

$$F\_measure = 2 \frac{Precision \cdot Recall}{Precision + Recall}$$

The similarity between two documents is determined based on the following equation:

$$Sim_{DocToDoc}(Q, C) = \frac{\sum_{q \in Q} Sim_{SentToDoc}(q, C)}{|Q|} \quad (3.10)$$

Where  $Q$  is a query document,  $C$  is a corpus/Web document,  $|Q|$  is the number of sentences in  $Q$  and  $Sim_{SentToDoc}(q, C)$  is a sentence-to-document similarity and given by the following equation

$$Sim_{SentToDoc}(q, C) = Max_{c \in C} \{sim(q, c) \geq t\} \quad (3.11)$$

Where  $q$  is a sentence in  $Q$ ,  $sim(q, c)$  is the similarity between sentence pairs as defined by either Equation 3.6, Figure 3.5, Figure 3.6, or Figure 3.7.

### **3.3 Summary**

This chapter briefly discussed the methodology that will be used in this project. Information about constructing the corpus, document preprocessing, and document representation were discussed. Similarity measures were also presented in the context of semantic networks and N-grams inverted index implementation. Finally a general framework for evaluating the findings was introduced.

## **CHAPTER 4**

### **EXPERIMENTAL RESULTS**

#### **4.1 Introduction**

This chapter presents the experimental results of this project obtained from 20 query document constructed manually from 10 web documents.

The first part of those results is a corpus-based and concerns about retrieving the original sentences. To ensure the validity of the semantic relatedness method in detecting most cases of plagiarism in English language, the results of this part were compared with N-grams using three well-known similarity measures.

The second part in the experiments focuses in retrieving the source documents from the Web. A proposed method based on named entities extraction shows that an exhaustive search is unnecessary and inappropriate for Web plagiarism detection.

Statistics about the corpus and information how the query documents were constructed are detailed in the next section.

## **4.2 Information about the Corpus**

10 Source documents were downloaded from ScienceDirect.com, the URLs, titles, and their corresponding categories are listed in Appendix B. From the source documents, 20 query documents were plagiarized using all the stated instances in section 3.2.2. Appendix E contains examples of about 50 original sentences and their plagiarized versions. The URLs of the 600 Wikipedia documents together with their corresponding categories are listed in Appendix C.

Details about the distribution of sentences over the query documents and information about their corresponding sources are listed in Table 4.1.

**Table 4.1** Number of plagiarized sentences in documents pairs (query-vertical)/(source -horizontal)

Document ID	1	2	3	4	5	6	7	8	9	10
#Sentences	82	185	107	373	367	168	206	106	281	260
1	13									
2	35	35								
3	33		33							
4	27			27						
5	32				32					
6	26					26				
7	37						37			
8	33							33		
9	40								40	
10	32									32
11	40								40	
12	27			27						
13	37						37			
14	32				32					
15	26					26				
16	25	5		6	9	5				
17	28	5		11	12					
18	21			8	7	6				
19	37			7			5		17	8
20	20	5		7	8					

Statistics about the query documents and documents in the corpus are shown in Table 4.2. In that table the number of valid sentences denotes those sentences that have at least 3 non-stop words. Sentences that do not satisfy this criterion were not considered in the experiments. The tokenized words are those words that are taken from the English alphabet (i.e., by removing punctuations, numbers, and other non essential tokens)

Statistics about documents' part-of-speech tagging are shown in Table 4.3. In this study the tagging was implemented using the Stanford POS Tagger[61]. The tagger is based on Maximum-Entropy model (CMM) and uses the Penn Treebank English POS tag set[62]. There are 44 tags in this set. Some of those tags are mapped to the basic part-of-speech tags that are used in WordNet (noun, verbs,

adjectives and adverbs) and the rest of them are discarded. Details about the tags' names, abbreviations, mapping, are shown in Appendix D.

**Table 4.2** Statistics about the corpus and query documents

	Query	Source	Corpus
Number of Documents	20	10	610
Size in Kilobyte	72.4 KB	306.75	15965.68
Number of Sentences	601	2236	116597
Number of valid sentences	601	2135	114304
Average sentence length	11	13	13
Number of tokenized words	10843	46216	2595447
Number of distinct words	2369	5216	77089
Number of distinct non-stop words	2292	5118	76980
Number of distinct non-stop ,stemmed words	1611	3408	56977

**Table 4.3** Statistics about part-of-speech tagging

	Query	Source	Corpus(first 110 documents)
Number of nouns	3738	16036	137695
Number of distinct nouns	1189	2806	17889
Number of verbs	1221	5236	44790
Number of distinct verbs	634	1474	6241
Number of adjectives	1094	4201	36686
Number of distinct adjectives	428	937	4821
Number of adverbs	262	1031	12051
Number of distinct adverbs	112	240	748

### 4.3 Sentence-to-Sentence similarity

This section details how the similarity between two sentences (a query sentence and a corpus sentence) was computed. The procedure of preprocessing and representing sentences varies based on the applied method, thus they are discussed further in two separate subsections. In Both techniques (N-grams and semantic relatedness) an example is given and the procedure is decomposed into its basic steps until reaching the similarity between the two sentences.

Each sentence in a query document is compared with every sentence in a corpus and a record about the maximum sentence-to-document similarity is kept. For example, the record [503 1 216 4 0.2041] tells that when sentence number 4 in the first query document was compared with document number 503 in the corpus, the most similar sentence was number 216 and the similarity was 0.2041.

#### 4.3.1 N-grams Approach

The gram sizes that have been tested in this study were 1, 2, 3, and 4 grams. For each gram size the similarity was computed using three well-known similarity measures, namely Cosine, Jaccard, and Dice coefficients. The following example computes the similarity using 1-gram with Jaccard similarity.

The query document  $Q=\{\text{The Biology Manufacturing System (BMS) aims to deal with non-foreseen changes in manufacturing environment. It is based on the ideas inspired by biology, like self-organization, evolution, learning and adaptation.}\}$ .

The sentence that has to be compared is  $s = \text{"It is based on biology aspects"}$ .

**Step 1:** *Preprocessing, removing stop-words and stemming the remaining word*

There are two sentences in  $Q$  the result of applying this step yields the following two sentences:

$q_1 = \text{"biologi manufactur system bm aim deal non-foreseen chang manufactur environ"}$ .

$q_2 = \text{"base idea inspir biologi like self-organ evolut learn adapt"}$ .

**Step 2:** *Constructing the grams vocabulary:*

The vocabulary  $V$  of unigrams consists of all unique words in  $Q$  of size 1 after applying step 1, i.e.,  $V = \{\text{biologi, manufactur, system, bm, aim, deal, non-foreseen, chang, environ, base, idea, inspire, like, self-organ, evolut, learn, adapt}\}$

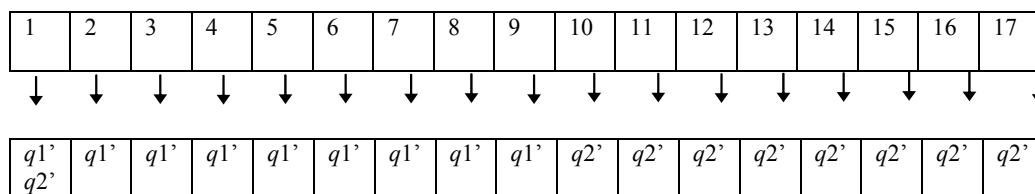
**Step 3:** *Generating the binary vectors for query sentences and constructing the inverted index*

Each sentence  $q$  is represented as a binary vector  $q'$  derived from  $V$  such that  $q'[i] = 1$  if  $V_i \in s$ , and  $q'[i] = 0$  otherwise.

Applying this to  $q_1$  and  $q_2$  yields the following two binary vectors;  
 $q_1'=[1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0]$  ,  $q_2'=[1,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1]$

The inverted index consists of a set of postings equals to the number of dimensions in the vocabulary, each post maps to a list of vectors that have non-zero entries in that dimension.

The inverted index for  $Q$  is a 17-diminsional index as depicted in Figure 4.1



**Figure 4.1** The inverted index for document  $Q$

**Step 4:** *computing the similarity*

Once the inverted index for the query document is built each sentence in the corpus is passed to step 1 and 3. In this example,  $s$  is converted to the following binary vector  $s'$

$$s'=[1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0]$$

Then the inverted index is scanned in a single pass to retrieve a list of candidate vectors from  $Q$ . In case of vector  $s'$ , both  $q1'$  and  $q2'$  are retrieved.

Finally the similarity is calculated using the three similarity measures - Jaccard, Dice, and Cosine - for the same sentence.

The Jaccard similarity between  $s$  and both  $q1$  and  $q2$  is:

$$J(s', q1') = \frac{1}{9 + 3 - 1} = 0.09$$

$$J(s', q2') = \frac{2}{9 + 3 - 2} = 0.20$$

Note that the size of the binary vector is the number of non-zero dimensions. For the corpus sentences, however, the size of the vector is replaced by the number of unique grams in the sentence. This was necessary since not all grams in the source sentence will be included in the vector representation.

### 4.3.2 Semantic Relatedness Approach

This section gives an example of the procedure of computing the semantic relatedness between two sentences based on equation 3.6.

Sentence 1= T1 =”It is based on the ideas inspired by biology, like self-organization, evolution, learning and adaptation.”Sentence 2= T2 =”It is based on biology aspects.”

The similarity between T1 and T2 equals 0.6683 and is obtained as follows:

**Step 1: part-of-speech tagging and preprocessing**

At first step the sentence is tagged in its original form, i.e., with all of its contents except punctuations. The reason is that the tagger needs all information about the sentence including functional words. Then all functional words such as conjunctions, prepositions, articles, auxiliary verbs, modal verbs, pronouns, cardinal numbers, and also punctuations are removed. The result of applying this step is shown in Table 4.4

**Table 4.4** Part-of-speech tagging of s1 and s2

word	Tag	Mapped		word	Tag	Mapped
It	PRP			it	PRP	
is	VBZ			is	VBZ	
based	VBN	Verb		based	VB	Verb
on	IN			on	IN	
the	DT			biolog	NN	Noun
ideas	NNS	Noun		aspect	NNS	Noun
inspired	VBN	Verb				
by	IN					
biology	NN	Noun				
like	IN					
Self-	NN	Noun				
evolution	NN	Noun				
learning	NN	Noun				
and	CC					
adaptation	NN	Noun				

**Step 2: creating the joint set**

The joint set contains all words in both sentences without duplicates. The joint set for *T1* and *T2* is:

Joint set= {based, ideas, inspired, biology, self-organization, evolution, learning, adaptation, aspects}.

**Step 3: Obtaining the semantic attributes:**

In this step WordNet is queried for each pair of words in the same part-of-speech . The semantic attributes between two words are the path between their synsets (synonym sets), and the depth of the synset that subsumes the two synsets (denoted the subsumer). For example, in Figure 4.2 the synset *{physical entity}* is a subsumer (and also *coordinate term*) of the two synsets *{process, physical process}* and *{object, physical object}*, the path between the two synsets is the number of nodes between the two synsets including the end node; which is 2 in this case, and the depth of this relation is the number of nodes along the hierarchy until reaching the topmost synset in the tree which equals 2 also (*{physical entity}* → *{entity}*). The overall procedure for obtaining the path between each part-of-speech word pair is defined by the procedure of Figure 3.2.

In many cases words in WordNet are polysemous, i.e., they have more than one sense. For example, in Figure 4.2 the noun *biology* has 3 senses. In such case only the shortest path is considered. Once shortest path is determined the subsumer depth is computed.

Figure 4.2 shows the hypernym trees of all nouns in sentence *T1* and the noun *biology* that exists in both sentences *T1* and *T2*. Figure 4.3 shows the hypernym trees of all nouns in sentence *T1* except the word *biology* and the hypernym trees of the second noun in *T2* *aspect*. Those trees are the actual trees in WordNet 2.1 for all senses.

Table 4.5 shows the shortest path between all pairs of nouns in the joint set and *T2*. Table 4.6 shows the corresponding relation depths, those attributes were obtained from Figure 4.2 and Figure 4.3.





**Table 4.5** Shortest path between word pairs in the joint set and  $T_2$  (“-1” no path exists, “=” equals, “?” not of the same part of speech)

	POS	Verb	Noun	Noun
POS	word	based	biology	aspects
Verb	based	=	?	?
Noun	ideas	?	7	4
Verb	inspired	-1	?	?
Noun	biology	?	=	7
Noun	self-organization	?	11	12
Noun	evolution	?	6	9
Noun	learning	?	8	6
Noun	adaptation	?	7	8
Noun	aspects	?	7	=

**Table 4.6** Subsumer depth between word pairs in the joint set and  $T_2$  (-1 no depth exists, = equals, ? not of the same part of speech)

	POS	Verb	Noun	Noun
POS	word	based	biology	aspects
Verb	based	=	?	?
Noun	ideas	?	6	7
Verb	inspired	-1	?	?
Noun	biology	?	=	3
Noun	self-organization	?	3	3
Noun	evolution	?	3	1
Noun	learning	?	3	6
Noun	adaptation	?	3	3
Noun	aspect	?	3	=

**Step 4:** *word-to-word similarity*

The similarity between words is a non-linear function of path length and depth and is defined by equation 3.1 which is repeated here for convenience:

$$s(w1, w2) = \begin{cases} 1, w1=w2 \\ 0, \text{no path exists or not of the same POS} \\ e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}}, \text{ otherwise} \end{cases}$$

Where  $l$  is the shortest path between  $w1$  and  $2$ ,  $h$  is the subsumer depth,  $\alpha$  scales the effect of path length and equals 0.2, and  $\beta$  scales the depth effect and equals 0.45. The result of word-to-word similarities for  $T2$  is shown in Table 4.7.

**Table 4.7** Word-to-word similarity between the joint set and  $T2$

	POS	Verb	Noun	Noun
POS	word	based	biology	aspects
Verb	based	1	0	0
Noun	ideas	0	0.2443	0.4476
Verb	inspired	0	0	0
Noun	biology	0	1	0.2155
Noun	self-organization	0	0.0968	0.0792
Noun	evolution	0	0.2632	0.0697
Noun	learning	0	0.1764	0.2984
Noun	adaptation	0	0.2155	0.1764
Noun	aspects	0	0.2155	1

**Step 5:** *deriving the raw semantic vectors and order vectors:*

The raw semantic vector length equals to the joint set size and its entry for a particular dimension equals to the maximum similarity at that dimension.

The order vector has the same properties of the semantic vector except that its entries are the relative positions of the maximum similar words in the sentence.

The similarity value between word-pairs must exceed the semantic and order thresholds to be considered in the raw semantic and order vectors respectively. The threshold is set to 0.2 in both cases.

Table 4.8 and Table 4.9 show the processes of deriving the raw semantic vectors and order vectors for  $T1$  and  $T2$  respectively.

**Table 4.8** Raw semantic and order vectors for  $T1$

	POS	Verb	Noun	Verb	Noun	Noun	Noun	Noun	Noun	Noun
POS	word	based	ideas	inspired	biology	self-organization	evolution	learning	adaptation	aspects
Verb	based	1	0	0	0	0	0	0	0	0
Noun	ideas	0	1	0	0.2443	0.1281	0.0697	0.5438	0.2334	0.4476
Verb	inspired	0	0	1	0	0	0	0	0	0
Noun	biology	0	0.2443	0	1	0.0968	0.2632	0.2000	0.2155	0.2155
Noun	self-organization	0	0.1281	0	0.0968	1	0.0313	0.1049	0.1341	0.0792
Noun	evolution	0	0.0697	0	0.2632	0.0313	1	0.0570	0.6346	0.0697
Noun	learning	0	0.5438	0	0.1764	0.1049	0.0570	1	0.1444	0.2984
Noun	adaptation	0	0.2334	0	0.2155	0.1341	0.6346	0.1444	1	0.1764

Thus for sentence  $T1$  the raw semantic vector=  
 $s1' = \{1,1,1,1,1,1,1,1,0.4476\}$  and the order vector= $r1 = \{1,2,3,4,5,6,7,8,2\}$

**Table 4.9** Raw semantic and order vectors for  $T2$ 

	POS	Verb	Noun	Verb	Noun	Noun	Noun	Noun	Noun	Noun
POS	word	based	ideas	inspired	biology	self-organization	evolution	learning	adaptation	aspects
Verb	based	1	0	0	0	0	0	0	0	0
Noun	biology	0	0.2443	0	1	0.0968	0.2632	0.2000	0.2155	0.2155
Noun	aspects	0	0.4476	0	0.2155	0.0792	0.0697	0.2984	0.1764	1

From Table 4.9 the raw semantic vector for  $T2 = s2' = \{1, 0.4476, 0, 1, 0, 0.2632, 0.2984, 0.2155, 1\}$  and the order vector  $r2 = \{1, 3, 0, 2, 2, 2, 3, 2, 3\}$ .

Note that in Table 4.9 the word *biology* was correctly mapped to the words *self-organization*, *evolution*, and *adaptation* since they are more semantically related to the noun *biology* than to the noun *aspect*. On the other hand the word *aspect* was automatically mapped to the words *ideas* and *learning*.

**Step 6:** calculating the Information Contents and obtaining the semantic vectors

The information content for a word  $w$  is derived from the Brown corpus as the probability of occurrence of that word in the corpus and given by equation 3.3:

$$I(w) = 1 - \frac{\log(n + 1)}{\log(N + 1)}$$

Where  $n$  is the number of occurrence of word  $w$  in the Brown corpus and  $N$  is the total number of words in the Brown corpus. There are 101, 594, 5 words in that corpus. In the experiments only the most frequent 5000 words in the Brown corpus are used, the list was obtained from [42]. The list constitutes 85% (865419 words) of the corpus and the minimum word occurrence in that list is 25.

A semantic vector entry,  $s_i$ , is given by equation 3.2:

$$s[i] = s'[i].I(wi).I(w'i)$$

Where  $wi$  is a word in the joint set and  $w'i$  is its associated word in the sentence. Table 4.10 shows the value  $n$  for each word in the joint set and its information content:

**Table 4.10** Information contents of word in the joint set

word	$n$	$I(w)$
based	119	0.6538
ideas	143	0.6406
inspired	25	0.7644
biology	0	1.0
self-organization	0	1.0
evolution	0	1.0
learning	60	0.7027
adaptation	0	1.0
aspects	64	0.6981

Hence the semantic vector for sentence 1 is : $s_1 = \{0.4275, 0.4105, 0.5844, 1, 1, 1, 0.4939, 1, 0.2003\}$

And the semantic vector for sentence 2 is:  $s_2 = \{0.4275, 0.2003, 0, 1, 0, 0.2633, 0.1465, 0.2155, 0.4875\}$

The semantic similarity  $S_s$  between sentence 1 and sentence 2 is given by the cosine coefficients (equation 3.4) between  $s_1$  and  $s_2$ , that is;

$$S_s = \frac{s_1 \cdot s_2}{\|s_1\| \cdot \|s_2\|} = 0.6786$$

**Step 7:** *The overall sentence to sentence similarity*

The order similarity  $S_r$  between sentence 1 and sentence 2 is obtained by the normalized difference of word order between the two sentences and given by equation 3.5:

$$S_r = 1 - \frac{\|r_1 - r_2\|}{\|r_1 + r_2\|} = 0.47241$$

Finally the overall similarity between sentence 1 and sentence 2 is given by equation 3.6

$$S(T_1, T_2) = \delta \cdot S_s + (1 - \delta) S_r$$

Where  $\delta$  decides the contribution of both semantic similarity,  $S_s$ , and order similarity,  $S_r$ . The value of  $\delta$  is set to 0.95. Hence;  $S(T_1, T_2) = 0.668353$

## 4.4 Results and Comparisons

The results of in this chapter are divided into two subsections, the first section is the corpus-based results and comparisons that mainly concerns about retrieving the original sentences from the corpus after applying the procedures of section 4.3. Section 4.4.2 shows and compares the results of applying three different techniques to retrieves the source documents from the Web. A full description about each technique is also included in that section.

### 4.4.1 Results of Corpus Sentence Retrieval

In most of the presented results in this section a 0.5 cutoff threshold was used. This is essential since all the query sentences are all plagiarized and should have a relatively high similarity score with the original sentences. Thus a 0.5 scoring threshold is fair to evaluate the performance of any method and have been used as a baseline to compare sentence similarity techniques [66]. Table 4.11 shows the recall rate of using N-grams with the three similarity measures when the 20 query documents were compared to the 610 document.

**Table 4.11** N-grams recall rate in 610 corpus documents with 0.5 cutoff threshold

	1-gram	2-grams	3-grams	4-grams
Cosine	0.6639	0.3344	0.2180	0.1381
Jaccard	0.4642	0.1930	0.1082	0.0965
Dice	0.6556	0.3311	0.2163	0.1364

Table 4.11 clearly shows that increasing the gram size reduces the recall rate significantly. This due to the fact that 2-grams (and in general any value of N greater than 1) will miss the original sentences in many cases. For example when the sentence is reordered without any change of its contents, 1-gram still gives a similarity equals to 1, however this is not necessary in 2, 3, or 4-grams. Also when a few words are replaced by synonyms any gram size greater than 1 will loss a large value of its similarity (depending on the gram size) since the comparison is between sentences and sentences are of small length.

The performance of Jaccard similarity was relatively poor comparing to Cosine and Dice coefficients in all gram sizes. In general, Cosine also slightly outperformed Dice coefficients in 2, 3, and 4-grams, thus in the following comparisons only the cosine similarity will be considered.

Table 4.12 shows the recall rate when comparing the 20 query documents against 110 documents in the corpus. The reason of using only 110 documents is that the semantic relatedness approach was computationally expensive.

To measure the increased number of false negatives when increasing the number of documents, the 20 query documents were compared with the 10 source documents, and the recall rate is obtained. Then another 10 documents from the corpus are added to the source documents and the recall rate is computed again. The process is repeated each time another 10 documents from the corpus were added until reaching 110 documents.

N-grams have reported a fewer false negatives than the semantic approach when increasing the number of documents (see the first 50, 60, and 70 documents in Table 4.12). Actually N-grams were consistent in terms of the recall rate until the 610 documents in the corpus except 1 gram which reported more false negatives than

2, 3, and 4-grams (compare the last row in Table 4.12 and the first row in Table 4.11).

**Table 4.12** Recall rate when increasing number of documents with 0.5 cutoff threshold

#Docs	#sentence pairs	cosine-1	cosine-2	cosine-3	cosine-4	Semantic R
10	601x2135	0.6656	0.3344	0.2180	0.1381	0.8419
20	601x3980	0.6656	0.3344	0.2180	0.1381	0.8369
30	601x5421	0.6656	0.3344	0.2180	0.1381	0.8369
40	601x7282	0.6656	0.3344	0.2180	0.1381	0.8369
50	601x9728	0.6656	0.3344	0.2180	0.1381	0.8369
60	601x11059	0.6656	0.3344	0.2180	0.1381	0.8353
70	601x12543	0.6656	0.3344	0.2180	0.1381	0.8336
80	601x14040	0.6656	0.3344	0.2180	0.1381	0.8336
90	601x16295	0.6656	0.3344	0.2180	0.1381	0.8336
100	601x17521	0.6656	0.3344	0.2180	0.1381	0.8336
110	601x19189	0.6656	0.3344	0.2180	0.1381	0.8336

Table 4.13 shows the number of true positives (TP), false positives (FP), false negatives (FN), precision, recall, and the harmonic mean in the 20x110 documents.

Out of the 601 query sentences, 1-gram cosine presented 5 false positives and one false positive in 2-grams. By manually comparing those 6 sentences with the original sentences none of them were more similar to the query sentences than the original sentences. 3 and 4-grams did not report any false positives (100% precision).

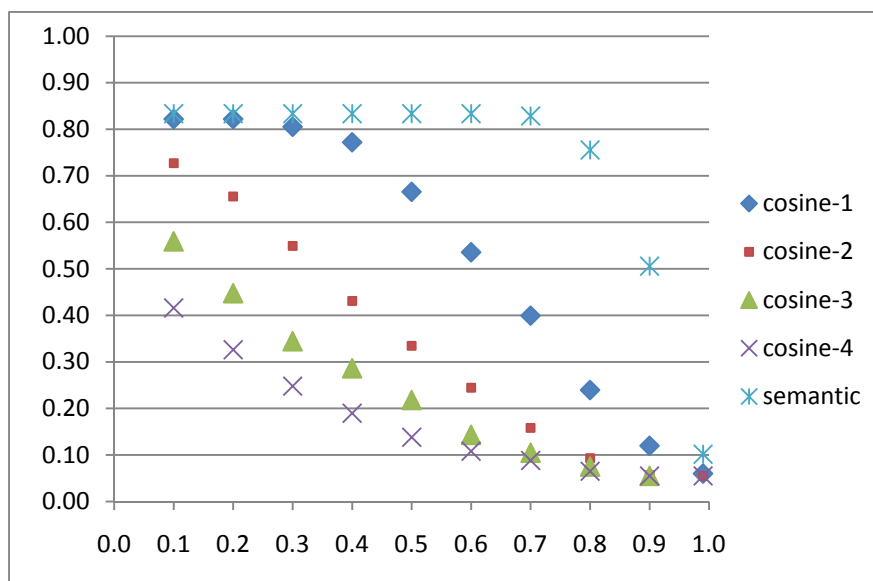
**Table 4.13** Precision, Recall, and Harmonic Mean (F-measure) in 110 corpus documents with 0.5 cutoff threshold

	TP	FP	FN	Precision	Recall	F-measure
cosine-1	400	5	201	0.9877	0.6656	0.7952
cosine-2	201	1	400	0.9950	0.3344	0.5006
cosine-3	131	0	470	1.0000	0.2180	0.3579
cosine-4	83	0	518	1.0000	0.1381	0.2427
Semantic R	507	95	94	0.8422	0.8436	0.8429

The F-measure in Table 4.13 shows that no major difference between semantic-R and cosine-1, however Table 4.13 shows only sentence pairs at 0.5 similarities. Since all query sentences are plagiarized, a 0.5 similarity is not sufficient enough to evaluate the performance of plagiarism detection methods and a good method is required to give a high similarity score. To further illustrate the accuracy of all methods in retrieving the original sentences, Table 4.14 shows the recall rate at all similarities ranges.

**Table 4.14** Recall rate across similarities in 110 corpus documents

Similarity >=	cosine-1	cosine-2	cosine-3	cosine-4	Semantic-R
0.1	0.8220	0.7271	0.5591	0.4160	0.8336
0.2	0.8220	0.6556	0.4476	0.3261	0.8336
0.3	0.8053	0.5491	0.3444	0.2479	0.8336
0.4	0.7720	0.4309	0.2862	0.1897	0.8336
0.5	0.6656	0.3344	0.2180	0.1381	0.8336
0.6	0.5358	0.2446	0.1431	0.1082	0.8336
0.7	0.3993	0.1581	0.1048	0.0882	0.8286
0.8	0.2396	0.0932	0.0749	0.0649	0.7554
0.9	0.1198	0.0566	0.0549	0.0549	0.5058
0.99	0.0599	0.0549	0.0549	0.0549	0.1015



**Figure 4.4** Recall rate (y-axis) across similarities (x-axis) in 110 corpus documents

Note that in Table 4.14 the recall rate of semantic approach was not affected by 0.5 similarity and still able to retrieve about 75% of the original sentences at 0.8 similarity comparing with 23% using 1-gram. This can be depicted in Figure 4.4.

Beyond 0.8 similarity the recall rate of the semantic approach starts to degrade significantly. At this point (0.8) the parameters of the algorithms were tested to get the optimal parameters. The algorithm depends on 5 parameters that contribute to the similarity between sentence pairs, namely; Alpha ( $\alpha$ ), Beta ( $\beta$ ), Delta ( $\delta$ ), the semantic threshold, and the order threshold in which their use have been presented in section 4.3.2. The values of Alpha (0.2) and Beta (0.45) and semantic threshold (0.2) have been optimized for WordNet in [57]. The value of Delta decides the contribution of both semantic and syntactic information between sentences. It has been shown that a similarity measure performs the best by giving the semantic information a higher weight than the syntactic information, in particular by setting this value to be higher than 80% [57]. Table 4.15 shows the recall rate by varying the value of Delta and the order threshold. The best recall rate was achieved by setting Delta to 0.95% and the order threshold to 0.1 (those are the values that were used in

all the comparisons in this section). This is intuitive in plagiarism detection since syntactic information is not important in measuring the similarity between sentences as many practices of plagiarism will involve changing the structure of sentences and order information.

**Table 4.15** Semantic-R recall rate in 110 corpus documents with Alpha=0.2, Beta=0.45 and 0.8 cutoff threshold

		Delta=	0.8	0.85	0.9	0.95
Order threshold	0.1		0.7105	0.7205	0.7354	0.7554
	0.2		0.7088	0.7188	0.7338	0.7537
	0.3		0.7038	0.7138	0.7338	0.7521
	0.4		0.6955	0.7121	0.7321	0.7504

#### 4.4.2 Results of Web Document Retrieval

This section presents the results of using the Google AJAX Web search API [65] in retrieving the 10 source documents from the Web. There are no restrictions in the API documentations on the number of allowed queries during some period of time. However Google limits the number of results that can be obtained from the API to 64 per query. Each search result consists of the title of the web document, its URL, and a short snippet that describe the web document.

For each query document a maximum number of 100 queries are extracted from that document and posted to the API. The returned URLs of each query are compared to the source URL(s) from which the query document originated to check

with a particular query was successful or not. The API was not directed to any site or domain in conducting these experiments.

There are three methods that have been used in the experiments. The first and the most basic one is by taking every three consecutive words (3-grams) within a sentence starting from the first sentence in the query document. Every 3-grams is then quoted (putted between quotations) to force Google to return the exact phrase. Table 4.16 shows the results of this approach. Starting from the left most column, the table shows the query document identifier, and for each document, the number of retrieved sources (note that in Table 4.1 some query documents were plagiarized from multiple sources), minimum number of queries required to retrieve that number of sources, number of missed sources, number of used queries by an algorithm, number of successful queries and finally the number of unique URLs that have been returned from all used queries. Stop words are not removed from 3-grams queries unless all three words were stop words, in such case the query is skipped and replaced by another one.

This searching scheme was not practical in many cases. Out of the 2000 queries that have been used only about 6% were successful. This small number of successful queries also entailed a large number of URLs and 11 sources were missed.

**Table 4.16** Results of using 3-grams searching with 64 results/query limit

	#found	within the 1st	#missed	#queries	#successful queries	#URLs
Doc ID						
1	1	23	0	100	7	2496
2	1	9	0	100	3	3379
3	1	14	0	100	2	3079
4	1	19	0	100	14	3155
5	1	6	0	100	8	3161
6	1	14	0	100	8	3261
7	1	13	0	100	5	3573
8	1	13	0	100	11	3429
9	1	34	0	100	10	3006
10	1	1	0	100	11	3060
11	1	1	0	100	3	2978
12	1	92	0	100	2	3202
13	1	87	0	100	2	3200
14	1	22	0	100	18	2852
15	1	51	0	100	6	3068
16	0		4	100	0	3384
17	2	89	1	100	5	3000
18	1	18	2	100	2	3102
19	1	3	3	100	2	2713
20	2	61	1	100	8	2721
<b>Total</b>	<b>21</b>		<b>11</b>	<b>2000</b>	<b>127</b>	<b>61819</b>

The second method is much like the previous one with a property of prioritizing the 3-grams queries by assigning a weight to each query. Queries are then ranked in a decreased order of their weights. The following is the weighting scheme that has been used for weighting a 3-grams  $G$ :

$$weight(G) = \sum_{w \in G} 1 - \frac{\log(wn + 1)}{\log(N + 1)}$$

Where  $w$  is a word in  $G$ , and as before  $wn$  is the number of occurrence of  $w$  in the Brown corpus,  $N$  is the total number of words in that corpus. As in the case of un-weighted 3-grams, a query that is completely stop words is not considered.

Table 4.17 shows the results of weighted 3-grams.

**Table 4.17** Results of using weighted 3-grams with 64 results/query limit

	#found	within the 1st	#missed	#queries	#successful queries	#URLs
Doc ID						
1	1	7	0	100	20	1515
2	1	1	0	100	38	1485
3	1	1	0	100	26	1412
4	1	1	0	100	42	2051
5	1	12	0	100	38	1777
6	1	1	0	100	22	2370
7	1	7	0	100	36	1618
8	1	1	0	100	32	2206
9	1	9	0	100	20	928
10	1	3	0	100	21	1861
11	1	12	0	100	10	833
12	1	2	0	100	25	2250
13	1	4	0	100	26	1733
14	1	4	0	100	37	1752
15	1	4	0	100	15	1809
16	0		4	100	0	1809
17	2	63	1	100	3	1983
18	1	52	2	100	2	2215
19	3	41	1	100	8	1180
20	2	10	1	100	11	1703
<b>Total</b>	<b>23</b>		<b>9</b>	<b>2000</b>	<b>432</b>	<b>34490</b>

The results in Table 4.17 are illustrative when compared to Table 4.16. For example, the percentage of successful queries has been increased to about 22%, the number of URLs have been decreased by an approximately 45%. Also the recall of retrieving the source documents was increased and the number of required queries to

retrieve the source documents was significantly reduced making this approach more attractive than the previous one.

However for some sophisticated instances of plagiarism this scheme fails in the retrieval process. For example, the 100 queries generated from document number 16 failed to retrieve one of its four sources. To overcome this limitation, another method was employed and is based on extracting named entities from sentences as the main primitive blocks of queries. Opposite to verbs, adjectives, adverbs and most common nouns, named entities such as proper nouns, names of agencies and locations are hard to be plagiarized. The extraction of named entities was implemented using the Stanford Named Entity Recognizer [67] (Stanford NER). The NER comes with two training models. One that trains the NER to classify entities into 3 classes (*person*, *location*, and *organization*) and the second and used model adds the *Misc* class to aforementioned classes. However named entities alone are not enough to construct queries in some cases such as where only one entity is present in a given sentence since it can be found in a large number of web documents. Thus the POS tagger is also used as a complementary tool to extract common nouns from the same sentence. The entities are quoted and placed in the left side of the query, and the remaining common nouns are quoted and placed in the right side of the query in a decreasing order of their importance according to equation 3.3. The query is then decomposed into multiple queries by reducing the number of quoted string once at a time until the number of quoted strings in the main query becomes 2. For illustration consider the following sentence:

*“Recently, the American National Institute of Building Sciences has inaugurated a committee to look into creating a standard for lifecycle data modelling under the BIM banner.”*

The NER classified the two following entities as organizations: *“American National Institute of Building Sciences”*, *“BIM”*.

The POS tagger further extracted the following five common nouns: “committee”, “standard”, “lifecycle”, “data”, “banner”. The followings are the generated 6 queries from the sentence after ordering the common nouns according to their weights:

*“American National Institute of Building Sciences”, “BIM”, “banner”, “committee”, “lifecycle”, “standard”, “data”*

*“American National Institute of Building Sciences”, “BIM”, “banner”, “committee”, “lifecycle”, “standard”*

*“American National Institute of Building Sciences”, “BIM”, “banner”, “committee”, “lifecycle”*

*“American National Institute of Building Sciences”, “BIM”, “banner”, “committee”*

*“American National Institute of Building Sciences”, “BIM”, “banner”*

*“American National Institute of Building Sciences”, “BIM”*

Table 4.18 shows the result of applying this selective searching. The number of successful queries increased to 33% comparing with 22% in weighted 3-grams and less number of URLs are returned. Note that even though the algorithm was allowed to use the 2000 query it required only 1109 query.

**Table 4.18** Results of using selective searching with 64 results/query limit

	#found	within the 1st	#missed	#queries	#successful queries	#URLs
Doc ID						
1	1	8	0	31	6	708
2	1	11	0	98	56	667
3	1	6	0	50	10	799
4	1	1	0	17	11	336
5	1	11	0	60	29	1191
6	1	1	0	3	1	89
7	1	1	0	100	44	2254
8	1	1	0	12	2	295
9	1	16	0	100	58	1756
10	1	3	0	100	46	1651
11	1	20	0	100	33	2016
12	1	8	0	12	4	375
13	1	2	0	79	16	2083
14	1	17	0	57	9	1109
15	0		1	10	0	7
16	2	29	2	88	17	1822
17	1	20	2	50	3	1014
18	1	5	2	13	4	408
19	2	67	2	79	3	1895
20	3	39	0	50	15	1138
<b>Total</b>	<b>23</b>		<b>9</b>	<b>1109</b>	<b>367</b>	<b>21613</b>

Tables 4.19 through 4.21 compare the three methods when the maximum results per query have been reduced to 8 results per query. In most cases the selective search outperforms the weighted and un-weighted 3-grams in several factors including the number of document have to be downloaded, the number of generated queries and the number of successful queries.

**Table 4.19** Results of using 3-grams searching with 8 results/query limit

	#found	within the 1st	#missed	#queries	#successful queries	#URLs
Doc ID						
1	1	23	0	100	6	336
2	1	9	0	100	2	441
3	1	14	0	100	1	424
4	1	19	0	100	9	421
5	1	81	0	100	5	431
6	1	47	0	100	4	425
7	1	13	0	100	3	459
8	1	13	0	100	6	456
9	1	34	0	100	9	401
10	1	19	0	100	3	412
11	1	68	0	100	2	395
12	1	92	0	100	2	429
13	1	87	0	100	2	440
14	1	37	0	100	10	408
15	1	51	0	100	6	408
16	0		4	100	0	440
17	1	89	2	100	3	430
18	1	18	2	100	1	412
19	1	29	3	100	1	375
20	1	19	2	100	1	374
<b>Total</b>	<b>19</b>		<b>13</b>	<b>2000</b>	<b>76</b>	<b>8317</b>

**Table 4.20** Results of using weighted 3-grams with 8 results/query limit

	#found	within the 1st	#missed	#queries	#successful queries	#URLs
Doc ID						
1	1	7	0	100	15	238
2	1	1	0	100	28	225
3	1	1	0	100	21	255
4	1	1	0	100	26	324
5	1	12	0	100	31	302
6	1	1	0	100	15	353
7	1	7	0	100	25	295
8	1	4	0	100	20	338
9	1	9	0	100	18	183
10	1	3	0	100	12	296
11	1	12	0	100	10	166
12	1	2	0	100	14	345
13	1	4	0	100	21	314
14	1	12	0	100	26	259
15	1	4	0	100	12	293
16	0		4	100	0	235
17	1	63	2	100	2	306
18	1	52	2	100	1	321
19	3	41	1	100	4	201
20	1	10	2	100	1	284
<b>Total</b>	<b>21</b>		<b>11</b>	<b>2000</b>	<b>302</b>	<b>5533</b>

**Table 4.21** Results of using selective searching with 8 results/query limit

	#found	within the 1st	#missed	#queries	#successful queries	#URLs
Doc ID						
1	1	8	0	31	4	96
2	1	11	0	98	52	129
3	1	29	0	50	8	117
4	1	10	0	17	7	49
5	1	35	0	60	18	155
6	0		1	3	0	9
7	1	1	0	100	40	276
8	1	1	0	12	2	43
9	1	16	0	100	50	244
10	1	6	0	100	31	222
11	1	20	0	99	28	257
12	1	8	0	12	1	54
13	1	2	0	79	12	260
14	1	38	0	57	6	144
15	0		1	10	0	4
16	2	42	2	88	13	257
17	1	20	2	50	2	130
18	1	5	2	13	1	55
19	1	12	3	79	1	259
20	1	1	2	50	6	153
<b>Total</b>	<b>19</b>		<b>13</b>	<b>1108</b>	<b>282</b>	<b>2913</b>

#### 4.4.3 Comparison with Existing Tools

The purpose of this section is twofold. First, to connect the finding presented in this chapter altogether in a basic application that takes a query document as an input (without any additional information) and outputs a ranked list of Web documents according to their similarity to the query document in a fully automated manner. Second, to evaluate the performance of the proposed method with respect to other freely available web-based tools.

The algorithm takes a query document and performs a selective search as discussed in section 4.4.2 followed by weighted 3-grams. Each search is limited to 100 query and 8 URLs per query. After the searching phase finishes, the URLs are simply ranked in a descending order by their frequencies. The most  $N$  frequent URLs for some given  $N$  are then selected and the corresponding documents are downloaded from the Web, parsed, and then saved locally. Equation 3.10 is then used to rank the downloaded documents. The value of  $t$  in equation 3.11 is set to 0.8.

For the best of our knowledge no freely available tool that had incorporated WordNet. Plagium [28] is the tool that was selected based on the initial experiment in section 2.4.

The framework of the evaluation is simple and straightforward. A plagiarized document is given to the algorithm and some tool and both are required to return the source document(s) in the top  $N$  of the ranked list. For example, if a query document was plagiarized from 3 web pages only the top 3 returned document are examined to check whether they contain the source documents and the recall rate is calculated. Formally the recall rate is defined as follows:

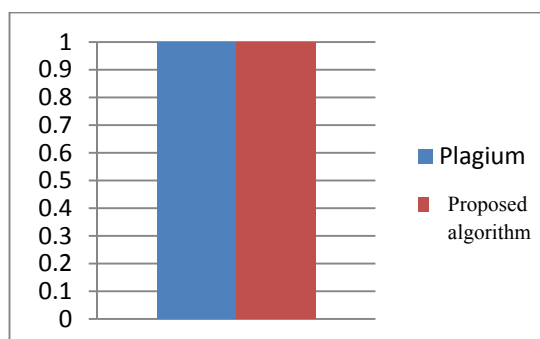
$$\text{recall rate} = \frac{\text{number of source documents in the top } N \text{ of the result}}{N}$$

Where  $N$  is the number of source documents.

Plagium and general other free tools have an access to digital libraries such as ScienceDirect. To test that the last 5 query documents used in the experiments were checked with Plagium and none of the source documents are returned either using the

raw or the plagiarized documents. Other documents were used in the comparison by entering two search keywords (“search engines” and “semantic relatedness”) to the ACM digital library (<http://portal.acm.org/dl.cfm>) and for each search, exactly five abstracts were selected such that Plagium can return the abstracts if they were entered in their original text. The ten abstracts were plagiarized by replacing words by their synonyms wherever it was possible but without affecting the mining of sentences and can be judged as plagiarized by a human inspector.

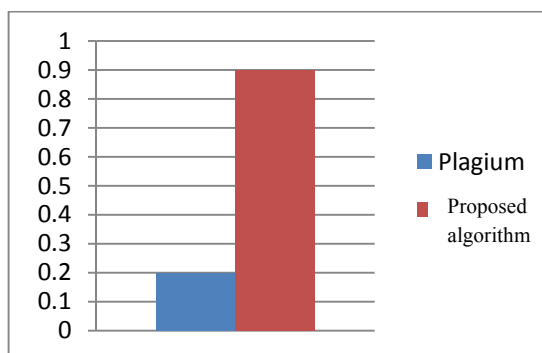
As before the Google API was not directed to ACM or any other domain. Figure 4.5 shows the recall rate of ten query documents each of which corresponds to one abstract as exact copy. The algorithm allowed to downloading only 10 documents for each query document. Both the algorithm and Plagium were able to retrieve the ten source documents.



**Figure 4.5** Recall rate in one-to-one exact copies

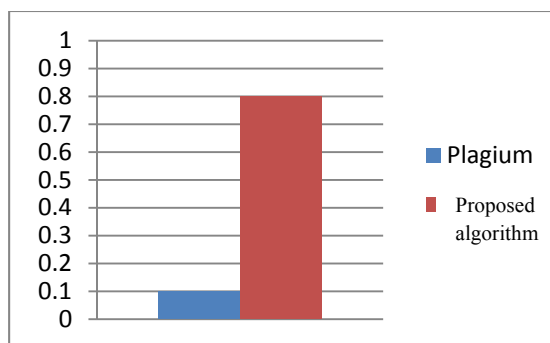
Figure 4.6 shows the recall rate of the plagiarized documents. As before the algorithm was limited to download 10 web documents per query document. Only one source document the algorithm had missed and the ranked list for that document was empty. The average similarity of equation 3.10 for the 9 documents was 0.64. Plagium returned only two source documents with average similarity equals to 0.28.

For the remaining 8 documents it showed a message indicating that the tool found no instances of plagiarism.



**Figure 4.6** Recall rate in one-to-one plagiarized by synonym replacing

Next a one-to-many test is carried out. Each five plagiarized abstracts that belong to the same search term (e.g., “search engines”) are grouped in one document. There are now two documents each of which plagiarized from 5 sources. Figure 4.7 shows the recall rate for this test. The algorithm was allowed to download 50 web documents per query document



**Figure 4.7** Recall rate in one-to-many plagiarized by synonym replacing

## 4.5 Discussion and Summary

The achieved findings in this chapter show that the semantic relatedness outperforms N-grams in most cases making this approach a valid methodology for detecting most cases of English language plagiarism. An important consideration is the number of false positives generated by this technique which was more than those found in N-grams. This comes from the fact that many words in WordNet are polysemous and presented in many concepts.

Obtaining the shortest path and ignoring word senses as it was applied in this project was a main contributor in those false matches. This false negatives problem was the main challenge in many semantic networks-based methods as reported in [51], nevertheless it can be reduced in the case of this project by integrating an appropriate word sense disambiguation functionality as discussed in Chapter 5.

A related aspect is the parameter setting of the algorithm. Although it is always subjective to human inspection, the results presented in Table 4.14 together with the example sentence pairs in Appendix E indicate that a 0.8 similarity score can be used as a cutoff point to highlight potential plagiarized sentences in most instances of plagiarism. The contribution of both semantic and syntactic information in computing the similarity between sentences can be concluded from Table 4.15. Table 4.15 shows that a higher contribution of the semantic and a low order threshold tend to increase the recall. However neglecting the order information of words within sentences completely (i.e., by setting the value Delta to 1) will treat sentences as bag-of-words which was not recommended in many literatures. Thus retaining a relatively small percentage (0.5% as indicated by Table 4.15) to the order information is required.

Another important remark comes from the web document retrieval tables. Note that in Table 4.21 the selective searching method missed the two sources of document number 6 and document number 15 whereas the weighted 3-grams searching (Table 4.20) efficiently retrieved the two sources of the two documents within the first 1 and 3 quires respectively. The reason is that in document number 6, the recognizer had found only one named entity in the document resulting in 3 queries only. The same problem presents in document number 15.

The selective searching based on named entities extraction has many promising properties including, reducing the search space, the high percentage of successful queries, and the most important property -from a web plagiarism detection perspective- the ability of retrieving the sources when a comprehensive plagiarism cases exist (see for example document number 16 in Table 4.21 and compare it with Table 4.20). However it has a drawback when named entities are not present in the plagiarized sentences, a problem which weighed grams do not suffer from. Thus it is useful to incorporate the weighted grams as a supplementary method for search expansion when only a few named entities are present in the suspected document. Alternatively grams that found to contain named entities can be combined in a proper context to constitute queries.

## **CHAPTER 5**

### **CONCLUSION**

#### **5.1 Introduction**

This study aimed to adopt semantic networks and general-purpose search engines for plagiarism detection in English documents. WordNet was the semantic network framework that has been used in this study, and the Google API was used in the experiments in retrieving the source documents from the Web.

The results in Chapter 4 show that the proposed algorithm was able to reveal most instances of natural language plagiarism and outperforms N-grams with different similarity measures. It also shows that retrieving the source documents from the Web is possible when some documents were moderately plagiarized. This chapter outlines the achievements, constraints, and future work of this study.

## 5.2 Achievements and Constraints

The achievements in this project can be outlined in the following remarks:

- Syntactic information alone not sufficient to reveal plagiarized sentences. This comes from the fact that order information is not important in computing the similarity between plagiarized sentences. A semantic relatedness between two sentences that is based on the path length of a semantic relation between their words, the depth of that relation, and information contents of words will increase the overall performance as recall gain will outweigh precision loss.
- Increasing the gram size in computing the similarity between short texts (e.g., sentences) that carry out plagiarism instances is not preferable. This is a consequence of lower recall rates when increasing the gram size. Additionally the percentage of precision loss is neglected when decreasing the gram size. The best performance can be achieved by unigrams.
- An exhaustive web searching using a search engine API is unnecessary to retrieve the source documents and has many drawbacks including a large list of documents to be downloaded, a small fraction of hits over misses. This can be avoided by extracting rare queries in a given corpus, or by extracting named entities and proper nouns since those are often hard to be plagiarized.

There were also some constraints in conducting the experiments. The main constraints can be identified in the following four points:

- Only some semantic relations were used with a focus on the “IS-A” and “synonymy” relations in obtaining the semantic attributes between word pairs.

- For words that found to be polysemous (have more than one sense) only the shortest path between word pairs is considered, regardless the actual senses.
- The comparison between words that are not within the same part-of-speech was limited to the equality.
- The behavior of the proposed algorithm was not assessed in sentence splitting/merging.

### 5.3 Future Work

The proposed framework in this study can be improved by including supports to different functionalities, for WordNet they include:

- *Word sense disambiguation*: An important functionality in measuring the semantic relatedness between word pairs in semantic networks. Neglecting polysemous words senses and taking the shortest path has a major disadvantage in that it could introduce false matching word pairs.
- *Utilizing other hierarchical relations*: Besides the “IS-A” (*hypernym/hyponym*) relation, it is widely accepted that other hierarchical relations such as the “HAS-A” (*holonym/meronym*) relation also contributes to the similarity between words and thus between sentences.
- *Using relations that cross part-of-speeches*: Although some of these relations were used in this project such as *pertain to* and *participle of*, they were used asymmetrically in adverbs and adjectives in cases

where no path exist between words within the same part-of-speech. There are also symmetric relations in WordNet that can be used to cross part-of-speeches such as nominalization.

- *Identifying the grammatical structure of sentences:* WordNet synsets also contain many collocations (e.g., “computer science”) that if they separated into single words would be found in different concepts. Thus identifying the grammatical structure of sentences is another important functionality. One method that often applied is by using natural language parsers.
- *Handling different inflected forms of words:* By stemming both words in sentences and WordNet, or more preferably by lemmatizing words in sentences; that is reducing words to a dictionary form so that they can still be found in WordNet. Alternatively, by keeping both original (for concept expansion) and inflected forms (for the actual computation) to use in a proper context.

For web document retrieval, future work might include one of the following techniques to reduce number of queries while maintaining an acceptable recall rate:

- Using information from the Web about the likelihood that two words in a query are similar to construct meaningful queries in order to avoid exhaustive searches. An example of such techniques is the Normalized Google Distance (NGD) [68], which measures the similarity between two words by using information about the number of Web pages that contain the two words separately and the number of pages that contain both words.
- *Using Document Summarization methods:* to filter out redundant information in the query document.

- *Using Stylometry Analysis methods:* to identify inconsistent writing styles within the query document in order to generate candidate queries.

## 5.4 Summary

An algorithm for document plagiarism detection using semantic networks has been proposed. Experimental results show that the algorithm was able to identify most of the plagiarized sentences in a high similarity range. The results also show that extracting named entities and nouns from sentences or ordering 3-grams queries based on their importance achieved promising recall rate in retrieving the source documents from the Web, even though only a few number sentences were plagiarized from the source documents. The performance of the presented techniques in this study can be further improved by several methods as briefly outlined in this chapter.

## REFERENCES

1. Lancaster, T., Culwin, F. *Classification of Plagiarism Detection Engines*. E-journal ITALICS, vol. 4 issue 2, ISSN 1473-7507., 2005
2. L. Huang. *A survey on web information retrieval technologies*. Computer Science. Dept., State Univ. New York, Stony Brook, NY, Tech. Rep., 2000.
3. Maurer, H., F. Kappe, B. Zaka. *Plagiarism – A Survey*. Journal of Universal Computer Sciences, vol. 12, no. 8, pp. 1050 – 1084, 2006.
4. <http://wordnet.princeton.edu/>.
5. C. Xiao, W. Wang, X. Lin, and J. X. Yu. *Efficient similarity joins for near duplicate detection*. In WWW, 2008.
6. R. J. Bayardo, Y. Ma, and R. Srikant. *Scaling up all pairs similarity search*. In WWW, 2007.
7. A. Arasu, V. Ganti, and R. Kaushik. *Efficient exact set-similarity joins*. In VLDB, 2006.
8. S. Sarawagi and A. Kirpal. *Efficient set joins on similarity predicates*. In SIGMOD, 2004.
9. Hassanzadeh, O., Sadoghi, M., and Miller, R. *Accuracy of Approximate String Joins Using Grams*. in VLDB, 2007.
10. SALTON, G., WONG, A., AND YANG, C. 1975. *A vector space model for automatic indexing*. Commun. ACM 18, 11, 613–620. Also reprinted in Sparck Jones and Willett [1997], pp. 273–280.
11. Goolge (<HTTP://www.Google.com>)
12. Gruner, S., S. Naven. *Tool support for plagiarism detection in text documents*. Proceedings of the 2005 ACM Symposium on Applied Computing. pp. 776 – 781, 2005.

13. Eissen, S., and Stein, B. *Intrinsic Plagiarism Detection*. Springer-Verlag ECIR LNCS 3936, pp. 565–569, 2006.
14. Manber, U. *Finding similar files in a large file system*. In Winter USENIX Technical Conference (pp. 1–10). San Francisco, CA., 1994.
15. Shivakumar, N., and Garcia-Molina, H. *Finding near-replicas of documents on the web*. In Proc. Workshop on Web Databases. 1998
16. Hoad, T. C. and Zobel, J. *Methods for Identifying Versioned and Plagiarised Documents*. Journal of the American Society for Information Science and Technology 54(3), 203–215. 2003.
17. Heintze, N. *Scalable document fingerprinting (extended abstract)*. In Proc. USENIX Workshop on Electronic Commerce. 1996.
18. S. Schleimer, DS Wilkerson, and A. Aiken. *Winnowing: local algorithms for document fingerprinting*. In Proceedings of the ACM SIGMOD International Conference on Management of Data. 2003
19. Shivakumar, N., and Garcia-Molina, H. *SCAM: a copy detection mechanism for digital documents*. In Proc. International Conference on Theory and Practice of Digital Libraries, Austin, Texas. 1995.
20. Brin, S., Davis, J., and Garcia-Molina, H. *Copy detection mechanisms for digital documents*. In Proc. ACM SIGMOD Annual Conference, San Jose, CA 1995.
21. Lyon, C., and Malcolm, J. *Demonstration of the Ferret Plagiarism Detector*. In Proceedings of the 2nd International Plagiarism Conference. 2006
22. Lyon, C., and Malcolm, J. *Detecting short passages of similar text in large document collections*. In Proceedings of Empirical Methods in Natural Language Processing Conference, pp. 118-125. 2001
23. Lyon C., Barrett R., and Malcolm J. *Plagiarism Is Easy, But Also Easy To Detect*. In Plagiary: Cross. Disciplinary Studies in Plagiarism. 2006
24. Broder., A. *On the resemblance and containment of documents*. In SEQS: Sequences '91, 1998.
25. Bao, J., Malcolm, J. *Text Similarity in Academic Conference Papers*. In: proceedings of the 2nd International Plagiarism Conference. The Sage, Gateshead. 2006.
26. Bao, J., Lyon, C. , Lane R., Ji, W., and Malcolm. J. *Copy detection in Chinese documents using the Ferret: A report on experiments*. Technical

- Report 456, Science and Technology Research Institute, University of Hertfordshire, 2006.
27. Barron-Cedeno, A., Rosso, P. *On Automatic Plagiarism Detection based on n-grams Comparison*. In: ECIR. LNCS, in press. 2009
  28. <http://www.plagium.com>
  29. Shivakumar, N., & Garcia-Molina, H. *Building a scalable and accurate copy detection mechanism*. In Proc. ACM Conference on Digital Libraries, Bethesda, MD. 1996
  30. S. Basu, R. J. Mooney, K. V. Pasupuleti, and J. Ghosh. *Evaluating the Novelty of Text-Mined Rules Using Lexical Knowledge*. In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001), 233-238, 2001.
  31. Ceska, Z, Toman, M., and Toman, K. *Multilingual Plagiarism Detection*. pringer-Verlag. AIMS 2008, LNAI 5253, pp. 83–92, 2008.
  32. Ceska, Z. *Plagiarism Detection Based on Singular Value Decomposition*. Springer-Verlag. LNAI 5221, pp. 108–119, 2008.
  33. R. Yerra and Y.-K. Ng. *A Sentence-Based Copy Detection Approach for Web Documents*. Proceedings of the 2nd Annual International Conference in Fuzzy Systems and Knowledge Discovery, pages 557-570. 2005.
  34. A. Moffat, J. Zobel. *Self-indexing inverted files for fast text retrieval*. ACM Trans. Inform. Syst. 14 (4) 349–379. 1996
  35. Zobel, J., Moffat, A., and Ramamohanarao, K. *Inverted files versus signature files for text indexing*. Technical Report CITRI/TR-95-5, Collaborative Information Technology Research Institute, Department of Computer Science, Royal Melbourne Institute of Technology, Australia, 1995.
  36. Kang, N., Gelbukh, A. *PPChecker: Plagiarism Pattern Checker in Document Copy Detection*. In: Sojka, P., Kopecek, I., Pala, K. (eds.) TSD 2006. LNCS, vol. 4188, pp. 661–667. Springer, Heidelberg. 2006
  37. Tachaphetpiboon, S., Facundes, N., and Amornraksa, T. *Plagiarism Indication by Syntactic-Semantic Analysis*. Proceedings of Asia-Pacific Conference on Communications 2007
  38. Clough, P. *Old and new challenges in automatic plagiarism detection*. Plagiarism Advisory Service, vol. 10, Department of Computer Science, University of Sheffield. 2003

39. C. Li, B. Wang, and X. Yang. *VGRAM: Improving performance of approximate queries on string collections using variable-length grams*. In VLDB, 2007.
40. Yi-Ting Liu, Heng-Rui Zhang, Tai-Wei Chen, and Wei-Guang Teng. *Extending Web Search for Online Plagiarism Detection*. IEEE 2007.
41. Takashi Tashiro, Takanori Ueda, Taisuke Hori, Yu Hirate and Hayato Yamana. *EPCI: Extracting Potentially Copyright Infringement Texts from the Web*. WWW 2007
42. <http://www.edict.com.hk/lexiconindex/frequencylists/>
43. <http://www.json.org/java/>
44. MONOSTORI, K., FINKEL, R. A., ZASLAVSKY, A., HODASZ, G., AND PATAKI, M. 2002. *Comparison of overlap detection techniques*. In International Conference on Computational Science. 2002.
45. Bao, J. Y. Shen, X. D. Liu, H. Y. Liu, and X. D. Zhang. *Semantic sequence kin: A method of document copy detection*. In Proceedings of the Advances in Knowledge Discovery and Data Mining, volume 3056, pages 529–538. Lecture Notes in Computer Science, 2004.
46. Bao, J. Y. Shen, X. D. Liu, H. Y. Liu, and X. D. Zhang. *Finding plagiarism based on common semantic sequence model*. In Proceedings of the 5th International Conference on Advances in Web-Age Information Management, volume 3129, pages 640–645. Lecture Notes in Computer Science, 2004.
47. Pataki, M. *Plagiarism Detection and Document Chunking Methods*. In WWW 2003
48. <http://www.doccop.com/>
49. Chaudhuri, S. Ganti, V. and Kaushik, R. *A primitive operator for similarity joins in data cleaning*. In Proc. of the 22nd Intl. Conf. on Data Engineering, 2006.
50. Sowa, J. F. (Eds.). (1992). *Principles of Semantic Networks*. San Mateo, CA: Morgan Kaufmann Publishers.
51. Alexander Budanitsky and Graeme Hirst. 2006. *Evaluating wordnet-based measures of lexical semantic relatedness*. Computational Linguistics, 32(1):13–47, 2006.

52. Graeme Hirst and David St-Onge. 1998. *Lexical chains as representations of context for the detection and correction of malapropisms*.
53. Sussna, Michael John. 1997. *Text Retrieval Using Inference in Semantic Metanetworks*. Ph.D. thesis, University of California, San Diego.
54. Wu, Zhibiao and Martha Palmer. 1994. *Verb semantics and lexical selection*. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, pages 133–138, Las Cruces, New Mexico, June.
55. Leacock, Claudia and Martin Chodorow. 1998. *Combining local context and WordNet similarity for word sense identification*. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA, chapter 11, pages 265–283.
56. P. Resnik, 1995. *Using Information Content to Evaluate Semantic Similarity in a Taxonomy*. Proc. 14th Int’l Joint Conf. AI.
57. Yuhua Li, David McLean, Zuhair A. Bandar, James D. O’Shea, and Keeley Crockett. 2006. *Sentence Similarity Based on Semantic Nets and Corpus Statistics*. IEEE Transactions on Knowledge and Data Engineering, 18(8):1138–1150.
58. Francis, Winthrop Nelson and Henry Kuřcera. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin, Boston.
59. <http://net.educause.edu/>
60. PORTER, M. 1980. *An algorithm for suffix stripping*. Program 14, 3, 130–137.
61. <http://nlp.stanford.edu/software/tagger.shtml>
62. Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. *Building a large annotated corpus of English: The Penn treebank*. Computational Linguistics 19 (2):313–330.
63. <http://www.ScienceDirect.com>
64. [http://en.wikipedia.org/wiki/Wikipedia:Featured\\_articles](http://en.wikipedia.org/wiki/Wikipedia:Featured_articles)
65. <http://code.google.com/apis/ajaxsearch/>
66. Achananuparp, P., Hu, X., Xiajiong, X. *The evaluation of sentence similarity measures*. In: Song, I.-Y., Eder, J., Nguyen, T.M. (eds.) DaWaK 2008. LNCS, vol. 5182, pp.305– 316. Springer, Heidelberg (2008).
67. <http://nlp.stanford.edu/software/CRF-NER.shtml>

68. Cilibrasi, R., Vitanyi, P. *The google similarity distance*. IEEE Transactions on knowledge and data engineering 19(3) (2007) 370–383
69. <http://sourceforge.net/projects/jwordnet/>
70. <http://www.turnitin.com/static/index.html>
71. <http://www.canexus.com/>

## APPENDIX A

### STOP-WORDS AND THEIR CORRESPONDING FREQUENCIES IN THE BROWN CORPUS

a	23363	inc	20	these	1573
about	1815	into	1791	they	3619
after	1070	is	10102	this	5146
all	3001	last	676	to	26154
also	1069	more	2216	up	1895
an	3748	most	1160	very	796
any	1345	mr	839	was	9815
and	28854	mrs	535	we	2653
are	4394	ms	null	well	897
as	7251	mz	null	were	3284
at	5377	no	2203	when	2331
be	6376	not	4610	where	938
because	883	only	1747	which	3561
been	2473	of	36410	who	2252
but	4381	on	6742	will	2244
by	5307	one	3297	with	7290
co	53	or	4207	would	2715
corp	null	other	1702		
could	1599	out	2096		
do	1362	over	1237		
for	9489	S	null		
from	4370	she	2859		
had	5131	so	1985		
has	2439	some	1617		
have	3942	say	504		
he	9542	says	200		
her	3037	such	1303		
his	6996	than	1790		
how	836	that	10594		
if	2199	the	69970		
it	8760	then	1377		
its	1858	their	2670		
in	21345	there	2725		

## APPENDIX B

### INFORMATION ABOUT SCIENCEDIRECT SOURCE DOCUMENTS

ID	Title	category	URL
1	Foldable subunits of helix protein	bioinformatics	<a href="http://dx.doi.org/10.1016/j.cmpbiolchem.2009.06.001">http://dx.doi.org/10.1016/j.cmpbiolchem.2009.06.001</a>
2	Computer integrated construction: A review and proposals for future direction	Advances in Engineering Software	<a href="http://dx.doi.org/10.1016/j.advengsoft.2006.10.007">http://dx.doi.org/10.1016/j.advengsoft.2006.10.007</a>
3	An imaging data model for concrete bridge inspection	Advances in Engineering Software	<a href="http://dx.doi.org/10.1016/j.advengsoft.2004.06.010">http://dx.doi.org/10.1016/j.advengsoft.2004.06.010</a>
4	A survey on real-world implementations of mobile ad-hoc networks	Ad Hoc Networks	<a href="http://dx.doi.org/10.1016/j.adhoc.2005.12.003">http://dx.doi.org/10.1016/j.adhoc.2005.12.003</a>
5	Evolutionary computing in manufacturing industry: an overview of recent applications	Applied Soft Computing	<a href="http://dx.doi.org/10.1016/j.asoc.2004.08.003">http://dx.doi.org/10.1016/j.asoc.2004.08.003</a>
6	Synthesis and emergence — research overview	Artificial Intelligence in Engineering	<a href="http://dx.doi.org/10.1016/S0954-1810(01)00022-X">http://dx.doi.org/10.1016/S0954-1810(01)00022-X</a>
7	Trends in network and service operation for the emerging future Internet Trends in network and service operation for the emerging future Internet	AEU - International Journal of Electronics and Communications	<a href="http://dx.doi.org/10.1016/j.aeu.2007.09.002">http://dx.doi.org/10.1016/j.aeu.2007.09.002</a>
8	Concept of self-reconfigurable modular robotic system	Artificial Intelligence in Engineering	<a href="http://dx.doi.org/10.1016/S0954-1810(01)00024-3">http://dx.doi.org/10.1016/S0954-1810(01)00024-3</a>
9	Enabling the creation of domain-specific reference collections to support text-based information retrievalnext term experiments in the architecture, engineering and construction industries	Advanced Engineering Informatics	<a href="http://dx.doi.org/10.1016/j.aei.2008.01.001">http://dx.doi.org/10.1016/j.aei.2008.01.001</a>
10	Applications of agent-based systems in intelligent manufacturing: An updated reviewnext term	Advanced Engineering Informatics	<a href="http://dx.doi.org/10.1016/j.aei.2006.05.004">http://dx.doi.org/10.1016/j.aei.2006.05.004</a>

## APPENDIX C

### INFORMATION ABOUT WIKIPEDIA CORPUS DOCUMENTS

ID	URL	Category
11	<a href="http://en.wikipedia.org/wiki/Azerbaijani_people">http://en.wikipedia.org/wiki/Azerbaijani_people</a>	
12	<a href="http://en.wikipedia.org/wiki/Daylight_saving_time">http://en.wikipedia.org/wiki/Daylight_saving_time</a>	
13	<a href="http://en.wikipedia.org/wiki/Turkey_Vulture">http://en.wikipedia.org/wiki/Turkey_Vulture</a>	Biology
14	<a href="http://en.wikipedia.org/wiki/Oceanic_whitetip_shark">http://en.wikipedia.org/wiki/Oceanic_whitetip_shark</a>	Biology
15	<a href="http://en.wikipedia.org/wiki/Immune_system">http://en.wikipedia.org/wiki/Immune_system</a>	Biology
16	<a href="http://en.wikipedia.org/wiki/California_Condor">http://en.wikipedia.org/wiki/California_Condor</a>	Biology
17	<a href="http://en.wikipedia.org/wiki/Australian_Green_Tree_Frog">http://en.wikipedia.org/wiki/Australian_Green_Tree_Frog</a>	Biology
18	<a href="http://en.wikipedia.org/wiki/Ocean_sunfish">http://en.wikipedia.org/wiki/Ocean_sunfish</a>	Biology
19	<a href="http://en.wikipedia.org/wiki/Guinea_pig">http://en.wikipedia.org/wiki/Guinea_pig</a>	Biology
20	<a href="http://en.wikipedia.org/wiki/Virus">http://en.wikipedia.org/wiki/Virus</a>	Biology
21	<a href="http://en.wikipedia.org/wiki/Peregrine_Falcon">http://en.wikipedia.org/wiki/Peregrine_Falcon</a>	Biology
22	<a href="http://en.wikipedia.org/wiki/Red-necked_Grebe">http://en.wikipedia.org/wiki/Red-necked_Grebe</a>	Biology
23	<a href="http://en.wikipedia.org/wiki/Red-tailed_Black_Cockatoo">http://en.wikipedia.org/wiki/Red-tailed_Black_Cockatoo</a>	Biology
24	<a href="http://en.wikipedia.org/wiki/Introduction_to_viruses">http://en.wikipedia.org/wiki/Introduction_to_viruses</a>	Biology
25	<a href="http://en.wikipedia.org/wiki/Island_Fox">http://en.wikipedia.org/wiki/Island_Fox</a>	Biology
26	<a href="http://en.wikipedia.org/wiki/Northern_Pintail">http://en.wikipedia.org/wiki/Northern_Pintail</a>	Biology
27	<a href="http://en.wikipedia.org/wiki/Sei_Whale">http://en.wikipedia.org/wiki/Sei_Whale</a>	Biology
28	<a href="http://en.wikipedia.org/wiki/Killer_Whale">http://en.wikipedia.org/wiki/Killer_Whale</a>	Biology
29	<a href="http://en.wikipedia.org/wiki/Parasaurolophus">http://en.wikipedia.org/wiki/Parasaurolophus</a>	Biology
30	<a href="http://en.wikipedia.org/wiki/Compsognathus">http://en.wikipedia.org/wiki/Compsognathus</a>	Biology
31	<a href="http://en.wikipedia.org/wiki/Jaguar">http://en.wikipedia.org/wiki/Jaguar</a>	Biology
32	<a href="http://en.wikipedia.org/wiki/Tarbosaurus">http://en.wikipedia.org/wiki/Tarbosaurus</a>	Biology
33	<a href="http://en.wikipedia.org/wiki/Right_whale">http://en.wikipedia.org/wiki/Right_whale</a>	Biology
34	<a href="http://en.wikipedia.org/wiki/Javan_Rhinoceros">http://en.wikipedia.org/wiki/Javan_Rhinoceros</a>	Biology
35	<a href="http://en.wikipedia.org/wiki/Chromatophore">http://en.wikipedia.org/wiki/Chromatophore</a>	Biology
36	<a href="http://en.wikipedia.org/wiki/Whale_song">http://en.wikipedia.org/wiki/Whale_song</a>	Biology
37	<a href="http://en.wikipedia.org/wiki/Sea_otter">http://en.wikipedia.org/wiki/Sea_otter</a>	Biology
38	<a href="http://en.wikipedia.org/wiki/Ant">http://en.wikipedia.org/wiki/Ant</a>	Biology
39	<a href="http://en.wikipedia.org/wiki/Komodo_dragon">http://en.wikipedia.org/wiki/Komodo_dragon</a>	Biology
40	<a href="http://en.wikipedia.org/wiki/Antbird">http://en.wikipedia.org/wiki/Antbird</a>	Biology
41	<a href="http://en.wikipedia.org/wiki/Cattle_Egret">http://en.wikipedia.org/wiki/Cattle_Egret</a>	Biology
42	<a href="http://en.wikipedia.org/wiki/Bird">http://en.wikipedia.org/wiki/Bird</a>	Biology
43	<a href="http://en.wikipedia.org/wiki/Procellariidae">http://en.wikipedia.org/wiki/Procellariidae</a>	Biology
44	<a href="http://en.wikipedia.org/wiki/Raccoon">http://en.wikipedia.org/wiki/Raccoon</a>	Biology
45	<a href="http://en.wikipedia.org/wiki/Cougar">http://en.wikipedia.org/wiki/Cougar</a>	Biology
46	<a href="http://en.wikipedia.org/wiki/Lion">http://en.wikipedia.org/wiki/Lion</a>	Biology
47	<a href="http://en.wikipedia.org/wiki/Blue_Whale">http://en.wikipedia.org/wiki/Blue_Whale</a>	Biology
48	<a href="http://en.wikipedia.org/wiki/Fin_Whale">http://en.wikipedia.org/wiki/Fin_Whale</a>	Biology

49	<a href="http://en.wikipedia.org/wiki/Humpback_Whale">http://en.wikipedia.org/wiki/Humpback_Whale</a>	Biology
50	<a href="http://en.wikipedia.org/wiki/Fauna_of_Scotland">http://en.wikipedia.org/wiki/Fauna_of_Scotland</a>	Biology
51	<a href="http://en.wikipedia.org/wiki/American_Black_Vulture">http://en.wikipedia.org/wiki/American_Black_Vulture</a>	Biology
52	<a href="http://en.wikipedia.org/wiki/Bacteria">http://en.wikipedia.org/wiki/Bacteria</a>	Biology
53	<a href="http://en.wikipedia.org/wiki/Emperor_Penguin">http://en.wikipedia.org/wiki/Emperor_Penguin</a>	Biology
54	<a href="http://en.wikipedia.org/wiki/Arctic_Tern">http://en.wikipedia.org/wiki/Arctic_Tern</a>	Biology
55	<a href="http://en.wikipedia.org/wiki/Cane_toad">http://en.wikipedia.org/wiki/Cane_toad</a>	Biology
56	<a href="http://en.wikipedia.org/wiki/Bald_Eagle">http://en.wikipedia.org/wiki/Bald_Eagle</a>	Biology
57	<a href="http://en.wikipedia.org/wiki/Banker_horse">http://en.wikipedia.org/wiki/Banker_horse</a>	Biology
58	<a href="http://en.wikipedia.org/wiki/Banksia_epica">http://en.wikipedia.org/wiki/Banksia_epica</a>	Biology
59	<a href="http://en.wikipedia.org/wiki/Banksia_spinulosa">http://en.wikipedia.org/wiki/Banksia_spinulosa</a>	Biology
60	<a href="http://en.wikipedia.org/wiki/Banksia_telmatiaea">http://en.wikipedia.org/wiki/Banksia_telmatiaea</a>	Biology
61	<a href="http://en.wikipedia.org/wiki/Blue_Iguana">http://en.wikipedia.org/wiki/Blue_Iguana</a>	Biology
62	<a href="http://en.wikipedia.org/wiki/Ficus_aurea">http://en.wikipedia.org/wiki/Ficus_aurea</a>	Biology
63	<a href="http://en.wikipedia.org/wiki/Alfred_Russel_Wallace">http://en.wikipedia.org/wiki/Alfred_Russel_Wallace</a>	Biology
64	<a href="http://en.wikipedia.org/wiki/Elfin-woods_Warbler">http://en.wikipedia.org/wiki/Elfin-woods_Warbler</a>	Biology
65	<a href="http://en.wikipedia.org/wiki/Red-backed_Fairy-wren">http://en.wikipedia.org/wiki/Red-backed_Fairy-wren</a>	Biology
66	<a href="http://en.wikipedia.org/wiki/Cyathus">http://en.wikipedia.org/wiki/Cyathus</a>	Biology
67	<a href="http://en.wikipedia.org/wiki/Cochineal">http://en.wikipedia.org/wiki/Cochineal</a>	Biology
68	<a href="http://en.wikipedia.org/wiki/Bobcat">http://en.wikipedia.org/wiki/Bobcat</a>	Biology
69	<a href="http://en.wikipedia.org/wiki/Amanita_muscaria">http://en.wikipedia.org/wiki/Amanita_muscaria</a>	Biology
70	<a href="http://en.wikipedia.org/wiki/Amanita_ocreata">http://en.wikipedia.org/wiki/Amanita_ocreata</a>	Biology
71	<a href="http://en.wikipedia.org/wiki/American_Goldfinch">http://en.wikipedia.org/wiki/American_Goldfinch</a>	Biology
72	<a href="http://en.wikipedia.org/wiki/Greater_Crested_Tern">http://en.wikipedia.org/wiki/Greater_Crested_Tern</a>	Biology
73	<a href="http://en.wikipedia.org/wiki/House_Martin">http://en.wikipedia.org/wiki/House_Martin</a>	Biology
74	<a href="http://en.wikipedia.org/wiki/Northern_Bald_Ibis">http://en.wikipedia.org/wiki/Northern_Bald_Ibis</a>	Biology
75	<a href="http://en.wikipedia.org/wiki/Seabird">http://en.wikipedia.org/wiki/Seabird</a>	Biology
76	<a href="http://en.wikipedia.org/wiki/Short-beaked_Echidna">http://en.wikipedia.org/wiki/Short-beaked_Echidna</a>	Biology
77	<a href="http://en.wikipedia.org/wiki/Shrimp_farm">http://en.wikipedia.org/wiki/Shrimp_farm</a>	Biology
78	<a href="http://en.wikipedia.org/wiki/Song_Thrush">http://en.wikipedia.org/wiki/Song_Thrush</a>	Biology
79	<a href="http://en.wikipedia.org/wiki/Elk">http://en.wikipedia.org/wiki/Elk</a>	Biology
80	<a href="http://en.wikipedia.org/wiki/Olm">http://en.wikipedia.org/wiki/Olm</a>	Biology
81	<a href="http://en.wikipedia.org/wiki/Proteasome">http://en.wikipedia.org/wiki/Proteasome</a>	Biology
82	<a href="http://en.wikipedia.org/wiki/Fauna_of_Puerto_Rico">http://en.wikipedia.org/wiki/Fauna_of_Puerto_Rico</a>	Biology
83	<a href="http://en.wikipedia.org/wiki/Fauna_of_Australia">http://en.wikipedia.org/wiki/Fauna_of_Australia</a>	Biology
84	<a href="http://en.wikipedia.org/wiki/Hawksbill_turtle">http://en.wikipedia.org/wiki/Hawksbill_turtle</a>	Biology
85	<a href="http://en.wikipedia.org/wiki/Platypus">http://en.wikipedia.org/wiki/Platypus</a>	Biology
86	<a href="http://en.wikipedia.org/wiki/Primate">http://en.wikipedia.org/wiki/Primate</a>	Biology
87	<a href="http://en.wikipedia.org/wiki/Kakapo">http://en.wikipedia.org/wiki/Kakapo</a>	Biology
88	<a href="http://en.wikipedia.org/wiki/Domestic_sheep">http://en.wikipedia.org/wiki/Domestic_sheep</a>	Biology
89	<a href="http://en.wikipedia.org/wiki/Phagocyte">http://en.wikipedia.org/wiki/Phagocyte</a>	Biology
90	<a href="http://en.wikipedia.org/wiki/King_Vulture">http://en.wikipedia.org/wiki/King_Vulture</a>	Biology
91	<a href="http://en.wikipedia.org/wiki/Knut_(polar_bear)">http://en.wikipedia.org/wiki/Knut_(polar_bear)</a>	Biology
92	<a href="http://en.wikipedia.org/wiki/Majungasaurus">http://en.wikipedia.org/wiki/Majungasaurus</a>	Biology
93	<a href="http://en.wikipedia.org/wiki/Myxobolus_cerebralis">http://en.wikipedia.org/wiki/Myxobolus_cerebralis</a>	Biology

94	<a href="http://en.wikipedia.org/wiki/Homo_floresiensis">http://en.wikipedia.org/wiki/Homo_floresiensis</a>	Biology
95	<a href="http://en.wikipedia.org/wiki/Mourning_Dove">http://en.wikipedia.org/wiki/Mourning_Dove</a>	Biology
96	<a href="http://en.wikipedia.org/wiki/Ediacara_biota">http://en.wikipedia.org/wiki/Ediacara_biota</a>	Biology
97	<a href="http://en.wikipedia.org/wiki/Suffolk_Punch">http://en.wikipedia.org/wiki/Suffolk_Punch</a>	Biology
98	<a href="http://en.wikipedia.org/wiki/Rufous-crowned_Sparrow">http://en.wikipedia.org/wiki/Rufous-crowned_Sparrow</a>	Biology
99	<a href="http://en.wikipedia.org/wiki/Stegosaurus">http://en.wikipedia.org/wiki/Stegosaurus</a>	Biology
100	<a href="http://en.wikipedia.org/wiki/Tawny_Owl">http://en.wikipedia.org/wiki/Tawny_Owl</a>	Biology
101	<a href="http://en.wikipedia.org/wiki/Tasmanian_Devil">http://en.wikipedia.org/wiki/Tasmanian_Devil</a>	Biology
102	<a href="http://en.wikipedia.org/wiki/Thoroughbred">http://en.wikipedia.org/wiki/Thoroughbred</a>	Biology
103	<a href="http://en.wikipedia.org/wiki/Thylacine">http://en.wikipedia.org/wiki/Thylacine</a>	Biology
104	<a href="http://en.wikipedia.org/wiki/Tree_Sparrow">http://en.wikipedia.org/wiki/Tree_Sparrow</a>	Biology
105	<a href="http://en.wikipedia.org/wiki/Edmontosaurus">http://en.wikipedia.org/wiki/Edmontosaurus</a>	Biology
106	<a href="http://en.wikipedia.org/wiki/Chiffchaff">http://en.wikipedia.org/wiki/Chiffchaff</a>	Biology
107	<a href="http://en.wikipedia.org/wiki/Albertosaurus">http://en.wikipedia.org/wiki/Albertosaurus</a>	Biology
108	<a href="http://en.wikipedia.org/wiki/Allosaurus">http://en.wikipedia.org/wiki/Allosaurus</a>	Biology
109	<a href="http://en.wikipedia.org/wiki/Nuthatch">http://en.wikipedia.org/wiki/Nuthatch</a>	Biology
110	<a href="http://en.wikipedia.org/wiki/Krill">http://en.wikipedia.org/wiki/Krill</a>	Biology
111	<a href="http://en.wikipedia.org/wiki/Lambeosaurus">http://en.wikipedia.org/wiki/Lambeosaurus</a>	Biology
112	<a href="http://en.wikipedia.org/wiki/Pinguicula_moranensis">http://en.wikipedia.org/wiki/Pinguicula_moranensis</a>	Biology
113	<a href="http://en.wikipedia.org/wiki/Flight_feather">http://en.wikipedia.org/wiki/Flight_feather</a>	Biology
114	<a href="http://en.wikipedia.org/wiki/Flocke">http://en.wikipedia.org/wiki/Flocke</a>	Biology
115	<a href="http://en.wikipedia.org/wiki/Georg_Forster">http://en.wikipedia.org/wiki/Georg_Forster</a>	Biology
116	<a href="http://en.wikipedia.org/wiki/Styracosaurus">http://en.wikipedia.org/wiki/Styracosaurus</a>	Biology
117	<a href="http://en.wikipedia.org/wiki/Superb_Fairy-wren">http://en.wikipedia.org/wiki/Superb_Fairy-wren</a>	Biology
118	<a href="http://en.wikipedia.org/wiki/Sumatran_Rhinoceros">http://en.wikipedia.org/wiki/Sumatran_Rhinoceros</a>	Biology
119	<a href="http://en.wikipedia.org/wiki/Common_Blackbird">http://en.wikipedia.org/wiki/Common_Blackbird</a>	Biology
120	<a href="http://en.wikipedia.org/wiki/Bone_Wars">http://en.wikipedia.org/wiki/Bone_Wars</a>	Biology
121	<a href="http://en.wikipedia.org/wiki/Common_Raven">http://en.wikipedia.org/wiki/Common_Raven</a>	Biology
122	<a href="http://en.wikipedia.org/wiki/Common_Treecreeper">http://en.wikipedia.org/wiki/Common_Treecreeper</a>	Biology
123	<a href="http://en.wikipedia.org/wiki/Velociraptor">http://en.wikipedia.org/wiki/Velociraptor</a>	Biology
124	<a href="http://en.wikipedia.org/wiki/Verbascum_thapsus">http://en.wikipedia.org/wiki/Verbascum_thapsus</a>	Biology
125	<a href="http://en.wikipedia.org/wiki/Willie_Wagtail">http://en.wikipedia.org/wiki/Willie_Wagtail</a>	Biology
126	<a href="http://en.wikipedia.org/wiki/Variegated_Fairy-wren">http://en.wikipedia.org/wiki/Variegated_Fairy-wren</a>	Biology
127	<a href="http://en.wikipedia.org/wiki/White-winged_Fairy-wren">http://en.wikipedia.org/wiki/White-winged_Fairy-wren</a>	Biology
128	<a href="http://en.wikipedia.org/wiki/Tyrannosaurus">http://en.wikipedia.org/wiki/Tyrannosaurus</a>	Biology
129	<a href="http://en.wikipedia.org/wiki/Amanita_phalloides">http://en.wikipedia.org/wiki/Amanita_phalloides</a>	Biology
130	<a href="http://en.wikipedia.org/wiki/Ailanthus_altissima">http://en.wikipedia.org/wiki/Ailanthus_altissima</a>	Biology
131	<a href="http://en.wikipedia.org/wiki/White-breasted_Nuthatch">http://en.wikipedia.org/wiki/White-breasted_Nuthatch</a>	Biology
132	<a href="http://en.wikipedia.org/wiki/G._Ledyard_Stebbins">http://en.wikipedia.org/wiki/G._Ledyard_Stebbins</a>	Biology
133	<a href="http://en.wikipedia.org/wiki/Thescelosaurus">http://en.wikipedia.org/wiki/Thescelosaurus</a>	Biology
134	<a href="http://en.wikipedia.org/wiki/Puerto_Rican_Amazon">http://en.wikipedia.org/wiki/Puerto_Rican_Amazon</a>	Biology
135	<a href="http://en.wikipedia.org/wiki/Ring-tailed_Lemur">http://en.wikipedia.org/wiki/Ring-tailed_Lemur</a>	Biology
136	<a href="http://en.wikipedia.org/wiki/Norman_Borlaug">http://en.wikipedia.org/wiki/Norman_Borlaug</a>	Biology
137	<a href="http://en.wikipedia.org/wiki/Andean_Condor">http://en.wikipedia.org/wiki/Andean_Condor</a>	Biology
138	<a href="http://en.wikipedia.org/wiki/Barn_Swallow">http://en.wikipedia.org/wiki/Barn_Swallow</a>	Biology

139	<a href="http://en.wikipedia.org/wiki/Pygmy_Hippopotamus">http://en.wikipedia.org/wiki/Pygmy_Hippopotamus</a>	Biology
140	<a href="http://en.wikipedia.org/wiki/Iguanodon">http://en.wikipedia.org/wiki/Iguanodon</a>	Biology
141	<a href="http://en.wikipedia.org/wiki/Chrysidia_rhipheus">http://en.wikipedia.org/wiki/Chrysidia_rhipheus</a>	Biology
142	<a href="http://en.wikipedia.org/wiki/Emu">http://en.wikipedia.org/wiki/Emu</a>	Biology
143	<a href="http://en.wikipedia.org/wiki/Gorgosaurus">http://en.wikipedia.org/wiki/Gorgosaurus</a>	Biology
144	<a href="http://en.wikipedia.org/wiki/Parallel_computing">http://en.wikipedia.org/wiki/Parallel_computing</a>	Computing
145	<a href="http://en.wikipedia.org/wiki/Search_engine_optimization">http://en.wikipedia.org/wiki/Search_engine_optimization</a>	Computing
146	<a href="http://en.wikipedia.org/wiki/The_Million_Dollar_Homepage">http://en.wikipedia.org/wiki/The_Million_Dollar_Homepage</a>	Computing
147	<a href="http://en.wikipedia.org/wiki/Microsoft">http://en.wikipedia.org/wiki/Microsoft</a>	Computing
148	<a href="http://en.wikipedia.org/wiki/Sequence_alignment">http://en.wikipedia.org/wiki/Sequence_alignment</a>	Computing
149	<a href="http://en.wikipedia.org/wiki/Macintosh">http://en.wikipedia.org/wiki/Macintosh</a>	Computing
150	<a href="http://en.wikipedia.org/wiki/35_mm_film">http://en.wikipedia.org/wiki/35_mm_film</a>	Engineering and technology
151	<a href="http://en.wikipedia.org/wiki/Archimedes">http://en.wikipedia.org/wiki/Archimedes</a>	Engineering and technology
152	<a href="http://en.wikipedia.org/wiki/Atomic_line_filter">http://en.wikipedia.org/wiki/Atomic_line_filter</a>	Engineering and technology
153	<a href="http://en.wikipedia.org/wiki/Autostereogram">http://en.wikipedia.org/wiki/Autostereogram</a>	Engineering and technology
154	<a href="http://en.wikipedia.org/wiki/Construction_of_the_World_Trade_Center">http://en.wikipedia.org/wiki/Construction_of_the_World_Trade_Center</a>	Engineering and technology
155	<a href="http://en.wikipedia.org/wiki/Caesar_cipher">http://en.wikipedia.org/wiki/Caesar_cipher</a>	Engineering and technology
156	<a href="http://en.wikipedia.org/wiki/Draining_and_development_of_the_Everglades">http://en.wikipedia.org/wiki/Draining_and_development_of_the_Everglades</a>	Engineering and technology
157	<a href="http://en.wikipedia.org/wiki/Electrical_engineering">http://en.wikipedia.org/wiki/Electrical_engineering</a>	Engineering and technology
158	<a href="http://en.wikipedia.org/wiki/Gas_metal_arc_welding">http://en.wikipedia.org/wiki/Gas_metal_arc_welding</a>	Engineering and technology
159	<a href="http://en.wikipedia.org/wiki/Gas_tungsten_arc_welding">http://en.wikipedia.org/wiki/Gas_tungsten_arc_welding</a>	Engineering and technology
160	<a href="http://en.wikipedia.org/wiki/Hanford_Site">http://en.wikipedia.org/wiki/Hanford_Site</a>	Engineering and technology
161	<a href="http://en.wikipedia.org/wiki/History_of_timekeeping_devices">http://en.wikipedia.org/wiki/History_of_timekeeping_devices</a>	Engineering and technology
162	<a href="http://en.wikipedia.org/wiki/Jarmann_M1884">http://en.wikipedia.org/wiki/Jarmann_M1884</a>	Engineering and technology
163	<a href="http://en.wikipedia.org/wiki/Kammerlader">http://en.wikipedia.org/wiki/Kammerlader</a>	Engineering and technology
164	<a href="http://en.wikipedia.org/wiki/Christopher_C._Kraft,_Jr.">http://en.wikipedia.org/wiki/Christopher_C._Kraft,_Jr.</a>	Engineering and technology
165	<a href="http://en.wikipedia.org/wiki/Krag-Petersson">http://en.wikipedia.org/wiki/Krag-Petersson</a>	Engineering and technology
166	<a href="http://en.wikipedia.org/wiki/Glynn_Lunney">http://en.wikipedia.org/wiki/Glynn_Lunney</a>	Engineering and technology
167	<a href="http://en.wikipedia.org/wiki/Panavision">http://en.wikipedia.org/wiki/Panavision</a>	Engineering and technology
168	<a href="http://en.wikipedia.org/wiki/Rampart_Dam">http://en.wikipedia.org/wiki/Rampart_Dam</a>	Engineering and technology
169	<a href="http://en.wikipedia.org/wiki/Renewable_energy_in_Scotland">http://en.wikipedia.org/wiki/Renewable_energy_in_Scotland</a>	Engineering and technology
170	<a href="http://en.wikipedia.org/wiki/Restoration_of_the_Everglades">http://en.wikipedia.org/wiki/Restoration_of_the_Everglades</a>	Engineering and technology
171	<a href="http://en.wikipedia.org/wiki/Scout_Moor_Wind_Farm">http://en.wikipedia.org/wiki/Scout_Moor_Wind_Farm</a>	Engineering and technology
172	<a href="http://en.wikipedia.org/wiki/Joseph_Francis_Shea">http://en.wikipedia.org/wiki/Joseph_Francis_Shea</a>	Engineering and technology
173	<a href="http://en.wikipedia.org/wiki/Shielded_metal_arc_welding">http://en.wikipedia.org/wiki/Shielded_metal_arc_welding</a>	Engineering and technology
174	<a href="http://en.wikipedia.org/wiki/Shuttle-Mir_Program">http://en.wikipedia.org/wiki/Shuttle-Mir_Program</a>	Engineering and technology
175	<a href="http://en.wikipedia.org/wiki/Space_Shuttle_Challenger_disaster">http://en.wikipedia.org/wiki/Space_Shuttle_Challenger_disaster</a>	Engineering and technology
176	<a href="http://en.wikipedia.org/wiki/Technology_of_the_Song_Dynasty">http://en.wikipedia.org/wiki/Technology_of_the_Song_Dynasty</a>	Engineering and technology
177	<a href="http://en.wikipedia.org/wiki/Welding">http://en.wikipedia.org/wiki/Welding</a>	Engineering and technology

178	<a href="http://en.wikipedia.org/wiki/World_Science_Festival">http://en.wikipedia.org/wiki/World_Science_Festival</a>	Engineering and technology
179	<a href="http://en.wikipedia.org/wiki/1928_Okeechobee_hurricane">http://en.wikipedia.org/wiki/1928_Okeechobee_hurricane</a>	Geology, geophysics and meteorology
180	<a href="http://en.wikipedia.org/wiki/1933_Atlantic_hurricane_season">http://en.wikipedia.org/wiki/1933_Atlantic_hurricane_season</a>	Geology, geophysics and meteorology
181	<a href="http://en.wikipedia.org/wiki/1980_eruption_of_Mount_St._Helens">http://en.wikipedia.org/wiki/1980_eruption_of_Mount_St._Helens</a>	Geology, geophysics and meteorology
182	<a href="http://en.wikipedia.org/wiki/1983_Atlantic_hurricane_season">http://en.wikipedia.org/wiki/1983_Atlantic_hurricane_season</a>	Geology, geophysics and meteorology
183	<a href="http://en.wikipedia.org/wiki/1988_Atlantic_hurricane_season">http://en.wikipedia.org/wiki/1988_Atlantic_hurricane_season</a>	Geology, geophysics and meteorology
184	<a href="http://en.wikipedia.org/wiki/1994_Atlantic_hurricane_season">http://en.wikipedia.org/wiki/1994_Atlantic_hurricane_season</a>	Geology, geophysics and meteorology
185	<a href="http://en.wikipedia.org/wiki/1995_Pacific_hurricane_season">http://en.wikipedia.org/wiki/1995_Pacific_hurricane_season</a>	Geology, geophysics and meteorology
186	<a href="http://en.wikipedia.org/wiki/1998_Pacific_hurricane_season">http://en.wikipedia.org/wiki/1998_Pacific_hurricane_season</a>	Geology, geophysics and meteorology
187	<a href="http://en.wikipedia.org/wiki/1999_Sydney_hailstorm">http://en.wikipedia.org/wiki/1999_Sydney_hailstorm</a>	Geology, geophysics and meteorology
188	<a href="http://en.wikipedia.org/wiki/2000_Sri_Lanka_cyclone">http://en.wikipedia.org/wiki/2000_Sri_Lanka_cyclone</a>	Geology, geophysics and meteorology
189	<a href="http://en.wikipedia.org/wiki/2002_Atlantic_hurricane_season">http://en.wikipedia.org/wiki/2002_Atlantic_hurricane_season</a>	Geology, geophysics and meteorology
190	<a href="http://en.wikipedia.org/wiki/2003_Atlantic_hurricane_season">http://en.wikipedia.org/wiki/2003_Atlantic_hurricane_season</a>	Geology, geophysics and meteorology
191	<a href="http://en.wikipedia.org/wiki/2005_Azores_subtropical_storm">http://en.wikipedia.org/wiki/2005_Azores_subtropical_storm</a>	Geology, geophysics and meteorology
192	<a href="http://en.wikipedia.org/wiki/2005_Atlantic_hurricane_season">http://en.wikipedia.org/wiki/2005_Atlantic_hurricane_season</a>	Geology, geophysics and meteorology
193	<a href="http://en.wikipedia.org/wiki/2006_Atlantic_hurricane_season">http://en.wikipedia.org/wiki/2006_Atlantic_hurricane_season</a>	Geology, geophysics and meteorology
194	<a href="http://en.wikipedia.org/wiki/2006_Pacific_hurricane_season">http://en.wikipedia.org/wiki/2006_Pacific_hurricane_season</a>	Geology, geophysics and meteorology
195	<a href="http://en.wikipedia.org/wiki/2007_Atlantic_hurricane_season">http://en.wikipedia.org/wiki/2007_Atlantic_hurricane_season</a>	Geology, geophysics and meteorology
196	<a href="http://en.wikipedia.org/wiki/Chicxulub_crater">http://en.wikipedia.org/wiki/Chicxulub_crater</a>	Geology, geophysics and meteorology
197	<a href="http://en.wikipedia.org/wiki/Climate_of_India">http://en.wikipedia.org/wiki/Climate_of_India</a>	Geology, geophysics and meteorology
198	<a href="http://en.wikipedia.org/wiki/Climate_of_Minnesota">http://en.wikipedia.org/wiki/Climate_of_Minnesota</a>	Geology, geophysics and meteorology
199	<a href="http://en.wikipedia.org/wiki/Eye_(cyclone)">http://en.wikipedia.org/wiki/Eye_(cyclone)</a>	Geology, geophysics and meteorology
200	<a href="http://en.wikipedia.org/wiki/Cyclone_Elita">http://en.wikipedia.org/wiki/Cyclone_Elita</a>	Geology, geophysics and meteorology
201	<a href="http://en.wikipedia.org/wiki/Effects_of_Hurricane_Isabel_in_Delaware">http://en.wikipedia.org/wiki/Effects_of_Hurricane_Isabel_in_Delaware</a>	Geology, geophysics and meteorology
202	<a href="http://en.wikipedia.org/wiki/Effects_of_Hurricane_Isabel_in_North_Carolina">http://en.wikipedia.org/wiki/Effects_of_Hurricane_Isabel_in_North_Carolina</a>	Geology, geophysics and meteorology
203	<a href="http://en.wikipedia.org/wiki/Effects_of_Hurricane_Ivan_in_the_Lesser_Antilles_and_South_America">http://en.wikipedia.org/wiki/Effects_of_Hurricane_Ivan_in_the_Lesser_Antilles_and_South_America</a>	Geology, geophysics and meteorology
204	<a href="http://en.wikipedia.org/wiki/Extratropical_cyclone">http://en.wikipedia.org/wiki/Extratropical_cyclone</a>	Geology, geophysics and meteorology
205	<a href="http://en.wikipedia.org/wiki/Acute_myeloid_leukemia">http://en.wikipedia.org/wiki/Acute_myeloid_leukemia</a>	Health and medicine
206	<a href="http://en.wikipedia.org/wiki/Alzheimer%27s_disease">http://en.wikipedia.org/wiki/Alzheimer%27s_disease</a>	Health and medicine
207	<a href="http://en.wikipedia.org/wiki/Anti-tobacco_movement_in_Nazi_Germany">http://en.wikipedia.org/wiki/Anti-tobacco_movement_in_Nazi_Germany</a>	Health and medicine
208	<a href="http://en.wikipedia.org/wiki/Asperger_syndrome">http://en.wikipedia.org/wiki/Asperger_syndrome</a>	Health and medicine
209	<a href="http://en.wikipedia.org/wiki/Autism">http://en.wikipedia.org/wiki/Autism</a>	Health and medicine
210	<a href="http://en.wikipedia.org/wiki/Frank_Macfarlane_Burnet">http://en.wikipedia.org/wiki/Frank_Macfarlane_Burnet</a>	Health and medicine
211	<a href="http://en.wikipedia.org/wiki/Helicobacter_pylori">http://en.wikipedia.org/wiki/Helicobacter_pylori</a>	Health and medicine
212	<a href="http://en.wikipedia.org/wiki/1960_South_Vietnamese_coup_attempt">http://en.wikipedia.org/wiki/1960_South_Vietnamese_coup_attempt</a>	History
213	<a href="http://en.wikipedia.org/wiki/1962_South_Vietnamese_Independence_Palace_bombing">http://en.wikipedia.org/wiki/1962_South_Vietnamese_Independence_Palace_bombing</a>	History
214	<a href="http://en.wikipedia.org/wiki/1964_Brinks_Hotel_bombing">http://en.wikipedia.org/wiki/1964_Brinks_Hotel_bombing</a>	History
215	<a href="http://en.wikipedia.org/wiki/1981_Irish_hunger_strike">http://en.wikipedia.org/wiki/1981_Irish_hunger_strike</a>	History
216	<a href="http://en.wikipedia.org/wiki/2007_Samjhauta_Express_bombings">http://en.wikipedia.org/wiki/2007_Samjhauta_Express_bombings</a>	History

217	<a href="http://en.wikipedia.org/wiki/Act_of_Independence_of_Lithuania">http://en.wikipedia.org/wiki/Act_of_Independence_of_Lithuania</a>	History
218	<a href="http://en.wikipedia.org/wiki/Samuel_Adams">http://en.wikipedia.org/wiki/Samuel_Adams</a>	History
219	<a href="http://en.wikipedia.org/wiki/Alcibiades">http://en.wikipedia.org/wiki/Alcibiades</a>	History
220	<a href="http://en.wikipedia.org/wiki/Ike_Altgens">http://en.wikipedia.org/wiki/Ike_Altgens</a>	History
221	<a href="http://en.wikipedia.org/wiki/Ancient_Egypt">http://en.wikipedia.org/wiki/Ancient_Egypt</a>	History
222	<a href="http://en.wikipedia.org/wiki/Anschluss">http://en.wikipedia.org/wiki/Anschluss</a>	History
223	<a href="http://en.wikipedia.org/wiki/Harriet_Arbuthnot">http://en.wikipedia.org/wiki/Harriet_Arbuthnot</a>	History
224	<a href="http://en.wikipedia.org/wiki/Arrest_and_assassination_of_Ngo_Dinh_Diem">http://en.wikipedia.org/wiki/Arrest_and_assassination_of_Ngo_Dinh_Diem</a>	History
225	<a href="http://en.wikipedia.org/wiki/Elias_Ashmole">http://en.wikipedia.org/wiki/Elias_Ashmole</a>	History
226	<a href="http://en.wikipedia.org/wiki/Aspasia">http://en.wikipedia.org/wiki/Aspasia</a>	History
227	<a href="http://en.wikipedia.org/wiki/Bath_School_disaster">http://en.wikipedia.org/wiki/Bath_School_disaster</a>	History
228	<a href="http://en.wikipedia.org/wiki/Ram%C3%B3n_Emeterio_Betances">http://en.wikipedia.org/wiki/Ram%C3%B3n_Emeterio_Betances</a>	History
229	<a href="http://en.wikipedia.org/wiki/Birmingham_campaign">http://en.wikipedia.org/wiki/Birmingham_campaign</a>	History
230	<a href="http://en.wikipedia.org/wiki/Stede_Bonnet">http://en.wikipedia.org/wiki/Stede_Bonnet</a>	History
231	<a href="http://en.wikipedia.org/wiki/Carsten_Borchgrevink">http://en.wikipedia.org/wiki/Carsten_Borchgrevink</a>	History
232	<a href="http://en.wikipedia.org/wiki/James_Bowie">http://en.wikipedia.org/wiki/James_Bowie</a>	History
233	<a href="http://en.wikipedia.org/wiki/Joel_Brand">http://en.wikipedia.org/wiki/Joel_Brand</a>	History
234	<a href="http://en.wikipedia.org/wiki/Isaac_Brock">http://en.wikipedia.org/wiki/Isaac_Brock</a>	History
235	<a href="http://en.wikipedia.org/wiki/Brown_Dog_affair">http://en.wikipedia.org/wiki/Brown_Dog_affair</a>	History
236	<a href="http://en.wikipedia.org/wiki/William_Speirs_Bruce">http://en.wikipedia.org/wiki/William_Speirs_Bruce</a>	History
237	<a href="http://en.wikipedia.org/wiki/Henry_Cornelius_Burnett">http://en.wikipedia.org/wiki/Henry_Cornelius_Burnett</a>	History
238	<a href="http://en.wikipedia.org/wiki/Byzantine_Empire">http://en.wikipedia.org/wiki/Byzantine_Empire</a>	History
239	<a href="http://en.wikipedia.org/wiki/California_Gold_Rush">http://en.wikipedia.org/wiki/California_Gold_Rush</a>	History
240	<a href="http://en.wikipedia.org/wiki/Chalukya_dynasty">http://en.wikipedia.org/wiki/Chalukya_dynasty</a>	History
241	<a href="http://en.wikipedia.org/wiki/Choe_Bu">http://en.wikipedia.org/wiki/Choe_Bu</a>	History
242	<a href="http://en.wikipedia.org/wiki/Chola_Dynasty">http://en.wikipedia.org/wiki/Chola_Dynasty</a>	History
243	<a href="http://en.wikipedia.org/wiki/William_Cooley">http://en.wikipedia.org/wiki/William_Cooley</a>	History
244	<a href="http://en.wikipedia.org/wiki/Confederate_government_of_Kentucky">http://en.wikipedia.org/wiki/Confederate_government_of_Kentucky</a>	History
245	<a href="http://en.wikipedia.org/wiki/Tom_Crean_(explorer)">http://en.wikipedia.org/wiki/Tom_Crean_(explorer)</a>	History
246	<a href="http://en.wikipedia.org/wiki/John_Dee">http://en.wikipedia.org/wiki/John_Dee</a>	History
247	<a href="http://en.wikipedia.org/wiki/Demosthenes">http://en.wikipedia.org/wiki/Demosthenes</a>	History
248	<a href="http://en.wikipedia.org/wiki/Discovery_Expedition">http://en.wikipedia.org/wiki/Discovery_Expedition</a>	History
249	<a href="http://en.wikipedia.org/wiki/Adriaen_van_der_Donck">http://en.wikipedia.org/wiki/Adriaen_van_der_Donck</a>	History
250	<a href="http://en.wikipedia.org/wiki/Double_Seven_Day_scuffle">http://en.wikipedia.org/wiki/Double_Seven_Day_scuffle</a>	History
251	<a href="http://en.wikipedia.org/wiki/Th%C3%ADch_Qu%E1%BA%A3ng_%C4%90%E1%BB%A9c">http://en.wikipedia.org/wiki/Th%C3%ADch_Qu%E1%BA%A3ng_%C4%90%E1%BB%A9c</a>	History
252	<a href="http://en.wikipedia.org/wiki/%C3%89cole_Polytechnique_massacre">http://en.wikipedia.org/wiki/%C3%89cole_Polytechnique_massacre</a>	History
253	<a href="http://en.wikipedia.org/wiki/Ehime_Maru_and_USS_Greenville_collision">http://en.wikipedia.org/wiki/Ehime_Maru_and_USS_Greenville_collision</a>	History
254	<a href="http://en.wikipedia.org/wiki/England_expects_that_every_man_will_do_his_duty">http://en.wikipedia.org/wiki/England_expects_that_every_man_will_do_his_duty</a>	History
255	<a href="http://en.wikipedia.org/wiki/Epaminondas">http://en.wikipedia.org/wiki/Epaminondas</a>	History
256	<a href="http://en.wikipedia.org/wiki/Anne_Frank">http://en.wikipedia.org/wiki/Anne_Frank</a>	History
257	<a href="http://en.wikipedia.org/wiki/French_Texas">http://en.wikipedia.org/wiki/French_Texas</a>	History
258	<a href="http://en.wikipedia.org/wiki/Mohandas_Karamchand_Gandhi">http://en.wikipedia.org/wiki/Mohandas_Karamchand_Gandhi</a>	History
259	<a href="http://en.wikipedia.org/wiki/Franklin_B._Gowen">http://en.wikipedia.org/wiki/Franklin_B._Gowen</a>	History
260	<a href="http://en.wikipedia.org/wiki/Gettysburg_Address">http://en.wikipedia.org/wiki/Gettysburg_Address</a>	History

261	<a href="http://en.wikipedia.org/wiki/Great_Fire_of_London">http://en.wikipedia.org/wiki/Great_Fire_of_London</a>	History
262	<a href="http://en.wikipedia.org/wiki/Hamlet_chicken_processing_plant_fire">http://en.wikipedia.org/wiki/Hamlet_chicken_processing_plant_fire</a>	History
263	<a href="http://en.wikipedia.org/wiki/Han_Dynasty">http://en.wikipedia.org/wiki/Han_Dynasty</a>	History
264	<a href="http://en.wikipedia.org/wiki/Richard_Hawes">http://en.wikipedia.org/wiki/Richard_Hawes</a>	History
265	<a href="http://en.wikipedia.org/wiki/Thomas_C._Hindman">http://en.wikipedia.org/wiki/Thomas_C._Hindman</a>	History
266	<a href="http://en.wikipedia.org/wiki/History_of_Arizona">http://en.wikipedia.org/wiki/History_of_Arizona</a>	History
267	<a href="http://en.wikipedia.org/wiki/History_of_the_Australian_Capital_Territory">http://en.wikipedia.org/wiki/History_of_the_Australian_Capital_Territory</a>	History
268	<a href="http://en.wikipedia.org/wiki/History_of_Burnside">http://en.wikipedia.org/wiki/History_of_Burnside</a>	History
269	<a href="http://en.wikipedia.org/wiki/History_of_the_Grand_Canyon_area">http://en.wikipedia.org/wiki/History_of_the_Grand_Canyon_area</a>	History
270	<a href="http://en.wikipedia.org/wiki/History_of_Lithuania_(1219%E2%80%931295)">http://en.wikipedia.org/wiki/History_of_Lithuania_(1219%E2%80%931295)</a>	History
271	<a href="http://en.wikipedia.org/wiki/History_of_Miami">http://en.wikipedia.org/wiki/History_of_Miami</a>	History
272	<a href="http://en.wikipedia.org/wiki/History_of_Minnesota">http://en.wikipedia.org/wiki/History_of_Minnesota</a>	History
273	<a href="http://en.wikipedia.org/wiki/History_of_New_Jersey">http://en.wikipedia.org/wiki/History_of_New_Jersey</a>	History
274	<a href="http://en.wikipedia.org/wiki/History_of_the_Philippines">http://en.wikipedia.org/wiki/History_of_the_Philippines</a>	History
275	<a href="http://en.wikipedia.org/wiki/History_of_Poland_(1945%E2%80%931989)">http://en.wikipedia.org/wiki/History_of_Poland_(1945%E2%80%931989)</a>	History
276	<a href="http://en.wikipedia.org/wiki/History_of_Portugal_(1777%E2%80%931834)">http://en.wikipedia.org/wiki/History_of_Portugal_(1777%E2%80%931834)</a>	History
277	<a href="http://en.wikipedia.org/wiki/History_of_Puerto_Rico">http://en.wikipedia.org/wiki/History_of_Puerto_Rico</a>	History
278	<a href="http://en.wikipedia.org/wiki/History_of_Tamil_Nadu">http://en.wikipedia.org/wiki/History_of_Tamil_Nadu</a>	History
279	<a href="http://en.wikipedia.org/wiki/History_of_Sheffield">http://en.wikipedia.org/wiki/History_of_Sheffield</a>	History
280	<a href="http://en.wikipedia.org/wiki/History_of_Solidarity">http://en.wikipedia.org/wiki/History_of_Solidarity</a>	History
281	<a href="http://en.wikipedia.org/wiki/History_of_the_Yosemite_area">http://en.wikipedia.org/wiki/History_of_the_Yosemite_area</a>	History
282	<a href="http://en.wikipedia.org/wiki/Hoysala_Empire">http://en.wikipedia.org/wiki/Hoysala_Empire</a>	History
283	<a href="http://en.wikipedia.org/wiki/Hungarian_Revolution_of_1956">http://en.wikipedia.org/wiki/Hungarian_Revolution_of_1956</a>	History
284	<a href="http://en.wikipedia.org/wiki/Imperial_Trans-Antarctic_Expedition">http://en.wikipedia.org/wiki/Imperial_Trans-Antarctic_Expedition</a>	History
285	<a href="http://en.wikipedia.org/wiki/Inaugural_games_of_the_Flavian_Amphitheatre">http://en.wikipedia.org/wiki/Inaugural_games_of_the_Flavian_Amphitheatre</a>	History
286	<a href="http://en.wikipedia.org/wiki/Jersey_Shore_shark_attacks_of_1916">http://en.wikipedia.org/wiki/Jersey_Shore_shark_attacks_of_1916</a>	History
287	<a href="http://en.wikipedia.org/wiki/Joan_of_Arc">http://en.wikipedia.org/wiki/Joan_of_Arc</a>	History
288	<a href="http://en.wikipedia.org/wiki/John_W._Johnston">http://en.wikipedia.org/wiki/John_W._Johnston</a>	History
289	<a href="http://en.wikipedia.org/wiki/Ernest_Joyce">http://en.wikipedia.org/wiki/Ernest_Joyce</a>	History
290	<a href="http://en.wikipedia.org/wiki/Katyn_massacre">http://en.wikipedia.org/wiki/Katyn_massacre</a>	History
291	<a href="http://en.wikipedia.org/wiki/Kengir_uprising">http://en.wikipedia.org/wiki/Kengir_uprising</a>	History
292	<a href="http://en.wikipedia.org/wiki/King_Arthur">http://en.wikipedia.org/wiki/King_Arthur</a>	History
293	<a href="http://en.wikipedia.org/wiki/Kingdom_of_Mysore">http://en.wikipedia.org/wiki/Kingdom_of_Mysore</a>	History
294	<a href="http://en.wikipedia.org/wiki/Shen_Kuo">http://en.wikipedia.org/wiki/Shen_Kuo</a>	History
295	<a href="http://en.wikipedia.org/wiki/Laika">http://en.wikipedia.org/wiki/Laika</a>	History
296	<a href="http://en.wikipedia.org/wiki/Lothal">http://en.wikipedia.org/wiki/Lothal</a>	History
297	<a href="http://en.wikipedia.org/wiki/Edward_Low">http://en.wikipedia.org/wiki/Edward_Low</a>	History
298	<a href="http://en.wikipedia.org/wiki/Aeneas_Mackintosh">http://en.wikipedia.org/wiki/Aeneas_Mackintosh</a>	History
299	<a href="http://en.wikipedia.org/wiki/Makuria">http://en.wikipedia.org/wiki/Makuria</a>	History
300	<a href="http://en.wikipedia.org/wiki/Charles_Edward_Magoon">http://en.wikipedia.org/wiki/Charles_Edward_Magoon</a>	History
301	<a href="http://en.wikipedia.org/wiki/Malcolm_X">http://en.wikipedia.org/wiki/Malcolm_X</a>	History
302	<a href="http://en.wikipedia.org/wiki/Manchester_Mummy">http://en.wikipedia.org/wiki/Manchester_Mummy</a>	History
303	<a href="http://en.wikipedia.org/wiki/Manzanar">http://en.wikipedia.org/wiki/Manzanar</a>	History
304	<a href="http://en.wikipedia.org/wiki/Marshall_Plan">http://en.wikipedia.org/wiki/Marshall_Plan</a>	History
305	<a href="http://en.wikipedia.org/wiki/Mauthausen-Gusen_concentration_camp">http://en.wikipedia.org/wiki/Mauthausen-Gusen_concentration_camp</a>	History

306	<a href="http://en.wikipedia.org/wiki/Harry_McNish">http://en.wikipedia.org/wiki/Harry_McNish</a>	History
307	<a href="http://en.wikipedia.org/wiki/Khalid_al-Mihdhar">http://en.wikipedia.org/wiki/Khalid_al-Mihdhar</a>	History
308	<a href="http://en.wikipedia.org/wiki/Ming_Dynasty">http://en.wikipedia.org/wiki/Ming_Dynasty</a>	History
309	<a href="http://en.wikipedia.org/wiki/Mormon_handcart_pioneers">http://en.wikipedia.org/wiki/Mormon_handcart_pioneers</a>	History
310	<a href="http://en.wikipedia.org/wiki/Benjamin_Morrell">http://en.wikipedia.org/wiki/Benjamin_Morrell</a>	History
311	<a href="http://en.wikipedia.org/wiki/Elizabeth_Needham">http://en.wikipedia.org/wiki/Elizabeth_Needham</a>	History
312	<a href="http://en.wikipedia.org/wiki/New_South_Greenland">http://en.wikipedia.org/wiki/New_South_Greenland</a>	History
313	<a href="http://en.wikipedia.org/wiki/Night_of_the_Long_Knives">http://en.wikipedia.org/wiki/Night_of_the_Long_Knives</a>	History
314	<a href="http://en.wikipedia.org/wiki/Nimrod_Expedition">http://en.wikipedia.org/wiki/Nimrod_Expedition</a>	History
315	<a href="http://en.wikipedia.org/wiki/Norte_Chico_civilization">http://en.wikipedia.org/wiki/Norte_Chico_civilization</a>	History
316	<a href="http://en.wikipedia.org/wiki/Emperor_Norton">http://en.wikipedia.org/wiki/Emperor_Norton</a>	History
317	<a href="http://en.wikipedia.org/wiki/Operation_Passage_to_Freedom">http://en.wikipedia.org/wiki/Operation_Passage_to_Freedom</a>	History
318	<a href="http://en.wikipedia.org/wiki/Rosa_Parks">http://en.wikipedia.org/wiki/Rosa_Parks</a>	History
319	<a href="http://en.wikipedia.org/wiki/Sardar_Vallabhbhai_Patel">http://en.wikipedia.org/wiki/Sardar_Vallabhbhai_Patel</a>	History
320	<a href="http://en.wikipedia.org/wiki/Pericles">http://en.wikipedia.org/wiki/Pericles</a>	History
321	<a href="http://en.wikipedia.org/wiki/Peterloo_Massacre">http://en.wikipedia.org/wiki/Peterloo_Massacre</a>	History
322	<a href="http://en.wikipedia.org/wiki/Rosa_Parks">http://en.wikipedia.org/wiki/Rosa_Parks</a>	History
323	<a href="http://en.wikipedia.org/wiki/Sardar_Vallabhbhai_Patel">http://en.wikipedia.org/wiki/Sardar_Vallabhbhai_Patel</a>	History
324	<a href="http://en.wikipedia.org/wiki/Pericles">http://en.wikipedia.org/wiki/Pericles</a>	History
325	<a href="http://en.wikipedia.org/wiki/Peterloo_Massacre">http://en.wikipedia.org/wiki/Peterloo_Massacre</a>	History
326	<a href="http://en.wikipedia.org/wiki/Phan_Dinh_Phung">http://en.wikipedia.org/wiki/Phan_Dinh_Phung</a>	History
327	<a href="http://en.wikipedia.org/wiki/Phan_Xich_Long">http://en.wikipedia.org/wiki/Phan_Xich_Long</a>	History
328	<a href="http://en.wikipedia.org/wiki/Witold_Pilecki">http://en.wikipedia.org/wiki/Witold_Pilecki</a>	History
329	<a href="http://en.wikipedia.org/wiki/Plymouth_Colony">http://en.wikipedia.org/wiki/Plymouth_Colony</a>	History
330	<a href="http://en.wikipedia.org/wiki/Polish%20%80%93Lithuanian_Commonwealth">http://en.wikipedia.org/wiki/Polish%20%80%93Lithuanian_Commonwealth</a>	History
331	<a href="http://en.wikipedia.org/wiki/Political_history_of_medieval_Karnataka">http://en.wikipedia.org/wiki/Political_history_of_medieval_Karnataka</a>	History
332	<a href="http://en.wikipedia.org/wiki/Political_integration_of_India">http://en.wikipedia.org/wiki/Political_integration_of_India</a>	History
333	<a href="http://en.wikipedia.org/wiki/Radhanite">http://en.wikipedia.org/wiki/Radhanite</a>	History
334	<a href="http://en.wikipedia.org/wiki/Sheikh_Mujibur_Rahman">http://en.wikipedia.org/wiki/Sheikh_Mujibur_Rahman</a>	History
335	<a href="http://en.wikipedia.org/wiki/Rashtrakuta_Dynasty">http://en.wikipedia.org/wiki/Rashtrakuta_Dynasty</a>	History
336	<a href="http://en.wikipedia.org/wiki/Red_Barn_Murder">http://en.wikipedia.org/wiki/Red_Barn_Murder</a>	History
337	<a href="http://en.wikipedia.org/wiki/Red_River_Trails">http://en.wikipedia.org/wiki/Red_River_Trails</a>	History
338	<a href="http://en.wikipedia.org/wiki/Retiarius">http://en.wikipedia.org/wiki/Retiarius</a>	History
339	<a href="http://en.wikipedia.org/wiki/Rock_Springs_massacre">http://en.wikipedia.org/wiki/Rock_Springs_massacre</a>	History
340	<a href="http://en.wikipedia.org/wiki/Woodes_Rogers">http://en.wikipedia.org/wiki/Woodes_Rogers</a>	History
341	<a href="http://en.wikipedia.org/wiki/Ross_Sea_party">http://en.wikipedia.org/wiki/Ross_Sea_party</a>	History
342	<a href="http://en.wikipedia.org/wiki/Rus%27_Khaganate">http://en.wikipedia.org/wiki/Rus%27_Khaganate</a>	History
343	<a href="http://en.wikipedia.org/wiki/S._A._Andr%C3%A9%27s_Arctic_balloon_expedition_of_1897">http://en.wikipedia.org/wiki/S._A._Andr%C3%A9%27s_Arctic_balloon_expedition_of_1897</a>	History
344	<a href="http://en.wikipedia.org/wiki/Saint-Sylvestre_coup_d%27%C3%A9tat">http://en.wikipedia.org/wiki/Saint-Sylvestre_coup_d%27%C3%A9tat</a>	History
345	<a href="http://en.wikipedia.org/wiki/Scotland_in_the_High_Middle_Ages">http://en.wikipedia.org/wiki/Scotland_in_the_High_Middle_Ages</a>	History
346	<a href="http://en.wikipedia.org/wiki/Robert_Falcon_Scott">http://en.wikipedia.org/wiki/Robert_Falcon_Scott</a>	History
347	<a href="http://en.wikipedia.org/wiki/Scottish_National_Antarctic_Expedition">http://en.wikipedia.org/wiki/Scottish_National_Antarctic_Expedition</a>	History
348	<a href="http://en.wikipedia.org/wiki/Second_Crusade">http://en.wikipedia.org/wiki/Second_Crusade</a>	History
349	<a href="http://en.wikipedia.org/wiki/Shackleton%20%80%93Rowett_Expedition">http://en.wikipedia.org/wiki/Shackleton%20%80%93Rowett_Expedition</a>	History

350	<a href="http://en.wikipedia.org/wiki/Ernest_Shackleton">http://en.wikipedia.org/wiki/Ernest_Shackleton</a>	History
351	<a href="http://en.wikipedia.org/wiki/Jack_Sheppard">http://en.wikipedia.org/wiki/Jack_Sheppard</a>	History
352	<a href="http://en.wikipedia.org/wiki/Wail_al-Shehri">http://en.wikipedia.org/wiki/Wail_al-Shehri</a>	History
353	<a href="http://en.wikipedia.org/wiki/Sino-German_cooperation_(1911%E2%80%931941)">http://en.wikipedia.org/wiki/Sino-German_cooperation_(1911%E2%80%931941)</a>	History
354	<a href="http://en.wikipedia.org/wiki/Slavery_in_ancient_Greece">http://en.wikipedia.org/wiki/Slavery_in_ancient_Greece</a>	History
355	<a href="http://en.wikipedia.org/wiki/Samantha_Smith">http://en.wikipedia.org/wiki/Samantha_Smith</a>	History
356	<a href="http://en.wikipedia.org/wiki/Song_Dynasty">http://en.wikipedia.org/wiki/Song_Dynasty</a>	History
357	<a href="http://en.wikipedia.org/wiki/Southern_Cross_Expedition">http://en.wikipedia.org/wiki/Southern_Cross_Expedition</a>	History
358	<a href="http://en.wikipedia.org/wiki/Suleiman_the_Magnificent">http://en.wikipedia.org/wiki/Suleiman_the_Magnificent</a>	History
359	<a href="http://en.wikipedia.org/wiki/Swedish_emigration_to_the_United_States">http://en.wikipedia.org/wiki/Swedish_emigration_to_the_United_States</a>	History
360	<a href="http://en.wikipedia.org/wiki/SY_Aurora%27s_drift">http://en.wikipedia.org/wiki/SY_Aurora%27s_drift</a>	History
361	<a href="http://en.wikipedia.org/wiki/Tang_Dynasty">http://en.wikipedia.org/wiki/Tang_Dynasty</a>	History
362	<a href="http://en.wikipedia.org/wiki/Terra_Nova_Expedition">http://en.wikipedia.org/wiki/Terra_Nova_Expedition</a>	History
363	<a href="http://en.wikipedia.org/wiki/Theramenes">http://en.wikipedia.org/wiki/Theramenes</a>	History
364	<a href="http://en.wikipedia.org/wiki/Tibet_during_the_Ming_Dynasty">http://en.wikipedia.org/wiki/Tibet_during_the_Ming_Dynasty</a>	History
365	<a href="http://en.wikipedia.org/wiki/To_the_People_of_Texas_%26_All_Americans_in_the_World">http://en.wikipedia.org/wiki/To_the_People_of_Texas_%26_All_Americans_in_the_World</a>	History
366	<a href="http://en.wikipedia.org/wiki/Treaty_of_Devol">http://en.wikipedia.org/wiki/Treaty_of_Devol</a>	History
367	<a href="http://en.wikipedia.org/wiki/Stephen_Trigg">http://en.wikipedia.org/wiki/Stephen_Trigg</a>	History
368	<a href="http://en.wikipedia.org/wiki/Hasekura_Tsunenaga">http://en.wikipedia.org/wiki/Hasekura_Tsunenaga</a>	History
369	<a href="http://en.wikipedia.org/wiki/Harriet_Tubman">http://en.wikipedia.org/wiki/Harriet_Tubman</a>	History
370	<a href="http://en.wikipedia.org/wiki/Vijayanagara_Empire">http://en.wikipedia.org/wiki/Vijayanagara_Empire</a>	History
371	<a href="http://en.wikipedia.org/wiki/Giovanni_Villani">http://en.wikipedia.org/wiki/Giovanni_Villani</a>	History
372	<a href="http://en.wikipedia.org/wiki/Voyage_of_the_James_Caird">http://en.wikipedia.org/wiki/Voyage_of_the_James_Caird</a>	History
373	<a href="http://en.wikipedia.org/wiki/Rudolf_Vrba">http://en.wikipedia.org/wiki/Rudolf_Vrba</a>	History
374	<a href="http://en.wikipedia.org/wiki/Roy_Welensky">http://en.wikipedia.org/wiki/Roy_Welensky</a>	History
375	<a href="http://en.wikipedia.org/wiki/Western_Chalukya_Empire">http://en.wikipedia.org/wiki/Western_Chalukya_Empire</a>	History
376	<a href="http://en.wikipedia.org/wiki/Western_Ganga_Dynasty">http://en.wikipedia.org/wiki/Western_Ganga_Dynasty</a>	History
377	<a href="http://en.wikipedia.org/wiki/Jonathan_Wild">http://en.wikipedia.org/wiki/Jonathan_Wild</a>	History
378	<a href="http://en.wikipedia.org/wiki/Yagan">http://en.wikipedia.org/wiki/Yagan</a>	History
379	<a href="http://en.wikipedia.org/wiki/Yellowstone_fires_of_1988">http://en.wikipedia.org/wiki/Yellowstone_fires_of_1988</a>	History
380	<a href="http://en.wikipedia.org/wiki/Zanzibar_Revolution">http://en.wikipedia.org/wiki/Zanzibar_Revolution</a>	History
381	<a href="http://en.wikipedia.org/wiki/Zhou_Tong_(archer)">http://en.wikipedia.org/wiki/Zhou_Tong_(archer)</a>	History
382	<a href="http://en.wikipedia.org/wiki/Ziad_Jarrah">http://en.wikipedia.org/wiki/Ziad_Jarrah</a>	History
383	<a href="http://en.wikipedia.org/wiki/Parapsychology">http://en.wikipedia.org/wiki/Parapsychology</a>	Philosophy and psychology
384	<a href="http://en.wikipedia.org/wiki/Conatus">http://en.wikipedia.org/wiki/Conatus</a>	Philosophy and psychology
385	<a href="http://en.wikipedia.org/wiki/S%C3%B8ren_Kierkegaard">http://en.wikipedia.org/wiki/S%C3%B8ren_Kierkegaard</a>	Philosophy and psychology
386	<a href="http://en.wikipedia.org/wiki/Eric_A._Havelock">http://en.wikipedia.org/wiki/Eric_A._Havelock</a>	Philosophy and psychology
387	<a href="http://en.wikipedia.org/wiki/Getting_It:_The_psychology_of_est">http://en.wikipedia.org/wiki/Getting_It:_The_psychology_of_est</a>	Philosophy and psychology
388	<a href="http://en.wikipedia.org/wiki/Bernard_Williams">http://en.wikipedia.org/wiki/Bernard_Williams</a>	Philosophy and psychology
389	<a href="http://en.wikipedia.org/wiki/Transhumanism">http://en.wikipedia.org/wiki/Transhumanism</a>	Philosophy and psychology
390	<a href="http://en.wikipedia.org/wiki/Hilary_Putnam">http://en.wikipedia.org/wiki/Hilary_Putnam</a>	Philosophy and psychology
391	<a href="http://en.wikipedia.org/wiki/Omnipotence_paradox">http://en.wikipedia.org/wiki/Omnipotence_paradox</a>	Philosophy and psychology
392	<a href="http://en.wikipedia.org/wiki/Philosophy_of_mind">http://en.wikipedia.org/wiki/Philosophy_of_mind</a>	Philosophy and psychology
393	<a href="http://en.wikipedia.org/wiki/Apollo_8">http://en.wikipedia.org/wiki/Apollo_8</a>	Physics and astronomy

394	<a href="http://en.wikipedia.org/wiki/Asteroid_belt">http://en.wikipedia.org/wiki/Asteroid_belt</a>	Physics and astronomy
395	<a href="http://en.wikipedia.org/wiki/Astrophysics_Data_System">http://en.wikipedia.org/wiki/Astrophysics_Data_System</a>	Physics and astronomy
396	<a href="http://en.wikipedia.org/wiki/Atmosphere_of_Jupiter">http://en.wikipedia.org/wiki/Atmosphere_of_Jupiter</a>	Physics and astronomy
397	<a href="http://en.wikipedia.org/wiki/Atom">http://en.wikipedia.org/wiki/Atom</a>	Physics and astronomy
398	<a href="http://en.wikipedia.org/wiki/Barnard%27s_Star">http://en.wikipedia.org/wiki/Barnard%27s_Star</a>	Physics and astronomy
399	<a href="http://en.wikipedia.org/wiki/Big_Bang">http://en.wikipedia.org/wiki/Big_Bang</a>	Physics and astronomy
400	<a href="http://en.wikipedia.org/wiki/Binary_star">http://en.wikipedia.org/wiki/Binary_star</a>	Physics and astronomy
401	<a href="http://en.wikipedia.org/wiki/Callisto_(moon)">http://en.wikipedia.org/wiki/Callisto_(moon)</a>	Physics and astronomy
402	<a href="http://en.wikipedia.org/wiki/Cat%27s_Eye_Nebula">http://en.wikipedia.org/wiki/Cat%27s_Eye_Nebula</a>	Physics and astronomy
403	<a href="http://en.wikipedia.org/wiki/Ceres_(dwarf_planet)">http://en.wikipedia.org/wiki/Ceres_(dwarf_planet)</a>	Physics and astronomy
404	<a href="http://en.wikipedia.org/wiki/Comet">http://en.wikipedia.org/wiki/Comet</a>	Physics and astronomy
405	<a href="http://en.wikipedia.org/wiki/Comet_Hale-Bopp">http://en.wikipedia.org/wiki/Comet_Hale-Bopp</a>	Physics and astronomy
406	<a href="http://en.wikipedia.org/wiki/Comet_Hyakutake">http://en.wikipedia.org/wiki/Comet_Hyakutake</a>	Physics and astronomy
407	<a href="http://en.wikipedia.org/wiki/Comet_Shoemaker-Levy_9">http://en.wikipedia.org/wiki/Comet_Shoemaker-Levy_9</a>	Physics and astronomy
408	<a href="http://en.wikipedia.org/wiki/Crab_Nebula">http://en.wikipedia.org/wiki/Crab_Nebula</a>	Physics and astronomy
409	<a href="http://en.wikipedia.org/wiki/Cygnus_X-1">http://en.wikipedia.org/wiki/Cygnus_X-1</a>	Physics and astronomy
410	<a href="http://en.wikipedia.org/wiki/Definition_of_planet">http://en.wikipedia.org/wiki/Definition_of_planet</a>	Physics and astronomy
411	<a href="http://en.wikipedia.org/wiki/Dwarf_planet">http://en.wikipedia.org/wiki/Dwarf_planet</a>	Physics and astronomy
412	<a href="http://en.wikipedia.org/wiki/Earth">http://en.wikipedia.org/wiki/Earth</a>	Physics and astronomy
413	<a href="http://en.wikipedia.org/wiki/Enceladus_(moon)">http://en.wikipedia.org/wiki/Enceladus_(moon)</a>	Physics and astronomy
414	<a href="http://en.wikipedia.org/wiki/Eris_(dwarf_planet)">http://en.wikipedia.org/wiki/Eris_(dwarf_planet)</a>	Physics and astronomy
415	<a href="http://en.wikipedia.org/wiki/Europa_(moon)">http://en.wikipedia.org/wiki/Europa_(moon)</a>	Physics and astronomy
416	<a href="http://en.wikipedia.org/wiki/Dwarf_planet">http://en.wikipedia.org/wiki/Dwarf_planet</a>	Physics and astronomy
417	<a href="http://en.wikipedia.org/wiki/Earth">http://en.wikipedia.org/wiki/Earth</a>	Physics and astronomy
418	<a href="http://en.wikipedia.org/wiki/Eris_(dwarf_planet)">http://en.wikipedia.org/wiki/Eris_(dwarf_planet)</a>	Physics and astronomy
419	<a href="http://en.wikipedia.org/wiki/Europa_(moon)">http://en.wikipedia.org/wiki/Europa_(moon)</a>	Physics and astronomy
420	<a href="http://en.wikipedia.org/wiki/Extrasolar_planet">http://en.wikipedia.org/wiki/Extrasolar_planet</a>	Physics and astronomy
421	<a href="http://en.wikipedia.org/wiki/Fermi_paradox">http://en.wikipedia.org/wiki/Fermi_paradox</a>	Physics and astronomy
422	<a href="http://en.wikipedia.org/wiki/Formation_and_evolution_of_the_Solar_System">http://en.wikipedia.org/wiki/Formation_and_evolution_of_the_Solar_System</a>	Physics and astronomy
423	<a href="http://en.wikipedia.org/wiki/Galaxy">http://en.wikipedia.org/wiki/Galaxy</a>	Physics and astronomy
424	<a href="http://en.wikipedia.org/wiki/Fermi_paradox">http://en.wikipedia.org/wiki/Fermi_paradox</a>	Physics and astronomy
425	<a href="http://en.wikipedia.org/wiki/Formation_and_evolution_of_the_Solar_System">http://en.wikipedia.org/wiki/Formation_and_evolution_of_the_Solar_System</a>	Physics and astronomy
426	<a href="http://en.wikipedia.org/wiki/Galaxy">http://en.wikipedia.org/wiki/Galaxy</a>	Physics and astronomy
427	<a href="http://en.wikipedia.org/wiki/Ganymede_(moon)">http://en.wikipedia.org/wiki/Ganymede_(moon)</a>	Physics and astronomy
428	<a href="http://en.wikipedia.org/wiki/General_relativity">http://en.wikipedia.org/wiki/General_relativity</a>	Physics and astronomy
429	<a href="http://en.wikipedia.org/wiki/Globular_cluster">http://en.wikipedia.org/wiki/Globular_cluster</a>	Physics and astronomy
430	<a href="http://en.wikipedia.org/wiki/H_II_region">http://en.wikipedia.org/wiki/H_II_region</a>	Physics and astronomy
431	<a href="http://en.wikipedia.org/wiki/GRB_970508">http://en.wikipedia.org/wiki/GRB_970508</a>	Physics and astronomy
432	<a href="http://en.wikipedia.org/wiki/Haumea_(dwarf_planet)">http://en.wikipedia.org/wiki/Haumea_(dwarf_planet)</a>	Physics and astronomy
433	<a href="http://en.wikipedia.org/wiki/Herbig%E2%80%93Haro_object">http://en.wikipedia.org/wiki/Herbig%E2%80%93Haro_object</a>	Physics and astronomy
434	<a href="http://en.wikipedia.org/wiki/Hubble_Deep_Field">http://en.wikipedia.org/wiki/Hubble_Deep_Field</a>	Physics and astronomy
435	<a href="http://en.wikipedia.org/wiki/Hubble_Space_Telescope">http://en.wikipedia.org/wiki/Hubble_Space_Telescope</a>	Physics and astronomy
436	<a href="http://en.wikipedia.org/wiki/IK_Pegasi">http://en.wikipedia.org/wiki/IK_Pegasi</a>	Physics and astronomy
437	<a href="http://en.wikipedia.org/wiki/Io_(moon)">http://en.wikipedia.org/wiki/Io_(moon)</a>	Physics and astronomy
438	<a href="http://en.wikipedia.org/wiki/Jupiter">http://en.wikipedia.org/wiki/Jupiter</a>	Physics and astronomy

439	<a href="http://en.wikipedia.org/wiki/Jupiter_Trojan">http://en.wikipedia.org/wiki/Jupiter_Trojan</a>	Physics and astronomy
440	<a href="http://en.wikipedia.org/wiki/Johannes_Kepler">http://en.wikipedia.org/wiki/Johannes_Kepler</a>	Physics and astronomy
441	<a href="http://en.wikipedia.org/wiki/Kreutz_Sungrazers">http://en.wikipedia.org/wiki/Kreutz_Sungrazers</a>	Physics and astronomy
442	<a href="http://en.wikipedia.org/wiki/Kuiper_belt">http://en.wikipedia.org/wiki/Kuiper_belt</a>	Physics and astronomy
443	<a href="http://en.wikipedia.org/wiki/Laplace%E2%80%93Runge%E2%80%93Lenz_vector">http://en.wikipedia.org/wiki/Laplace%E2%80%93Runge%E2%80%93Lenz_vector</a>	Physics and astronomy
444	<a href="http://en.wikipedia.org/wiki/Mars">http://en.wikipedia.org/wiki/Mars</a>	Physics and astronomy
445	<a href="http://en.wikipedia.org/wiki/Mercury_(planet)">http://en.wikipedia.org/wiki/Mercury_(planet)</a>	Physics and astronomy
446	<a href="http://en.wikipedia.org/wiki/Moon">http://en.wikipedia.org/wiki/Moon</a>	Physics and astronomy
447	<a href="http://en.wikipedia.org/wiki/Neptune">http://en.wikipedia.org/wiki/Neptune</a>	Physics and astronomy
448	<a href="http://en.wikipedia.org/wiki/Planet">http://en.wikipedia.org/wiki/Planet</a>	Physics and astronomy
449	<a href="http://en.wikipedia.org/wiki/Pluto">http://en.wikipedia.org/wiki/Pluto</a>	Physics and astronomy
450	<a href="http://en.wikipedia.org/wiki/Planets_beyond_Neptune">http://en.wikipedia.org/wiki/Planets_beyond_Neptune</a>	Physics and astronomy
451	<a href="http://en.wikipedia.org/wiki/Rings_of_Jupiter">http://en.wikipedia.org/wiki/Rings_of_Jupiter</a>	Physics and astronomy
452	<a href="http://en.wikipedia.org/wiki/Rings_of_Neptune">http://en.wikipedia.org/wiki/Rings_of_Neptune</a>	Physics and astronomy
453	<a href="http://en.wikipedia.org/wiki/Rings_of_Uranus">http://en.wikipedia.org/wiki/Rings_of_Uranus</a>	Physics and astronomy
454	<a href="http://en.wikipedia.org/wiki/Saturn">http://en.wikipedia.org/wiki/Saturn</a>	Physics and astronomy
455	<a href="http://en.wikipedia.org/wiki/Solar_eclipse">http://en.wikipedia.org/wiki/Solar_eclipse</a>	Physics and astronomy
456	<a href="http://en.wikipedia.org/wiki/Solar_System">http://en.wikipedia.org/wiki/Solar_System</a>	Physics and astronomy
457	<a href="http://en.wikipedia.org/wiki/Star">http://en.wikipedia.org/wiki/Star</a>	Physics and astronomy
458	<a href="http://en.wikipedia.org/wiki/Sun">http://en.wikipedia.org/wiki/Sun</a>	Physics and astronomy
459	<a href="http://en.wikipedia.org/wiki/Supernova">http://en.wikipedia.org/wiki/Supernova</a>	Physics and astronomy
460	<a href="http://en.wikipedia.org/wiki/Vega">http://en.wikipedia.org/wiki/Vega</a>	Physics and astronomy
461	<a href="http://en.wikipedia.org/wiki/Venus">http://en.wikipedia.org/wiki/Venus</a>	Physics and astronomy
462	<a href="http://en.wikipedia.org/wiki/1880_Republican_National_Convention">http://en.wikipedia.org/wiki/1880_Republican_National_Convention</a>	Politics and government
463	<a href="http://en.wikipedia.org/wiki/1996_United_States_campaign_finance_controversy">http://en.wikipedia.org/wiki/1996_United_States_campaign_finance_controversy</a>	Politics and government
464	<a href="http://en.wikipedia.org/wiki/Anarcho-capitalism">http://en.wikipedia.org/wiki/Anarcho-capitalism</a>	Politics and government
465	<a href="http://en.wikipedia.org/wiki/Yasser_Arafat">http://en.wikipedia.org/wiki/Yasser_Arafat</a>	Politics and government
466	<a href="http://en.wikipedia.org/wiki/Ban_Ki-moon">http://en.wikipedia.org/wiki/Ban_Ki-moon</a>	Politics and government
467	<a href="http://en.wikipedia.org/wiki/Alexandre_Banza">http://en.wikipedia.org/wiki/Alexandre_Banza</a>	Politics and government
468	<a href="http://en.wikipedia.org/wiki/Barth%C3%A9lemy_Boganda">http://en.wikipedia.org/wiki/Barth%C3%A9lemy_Boganda</a>	Politics and government
469	<a href="http://en.wikipedia.org/wiki/John_Brownlee_sex_scandal">http://en.wikipedia.org/wiki/John_Brownlee_sex_scandal</a>	Politics and government
470	<a href="http://en.wikipedia.org/wiki/Canadian_federal_election,_1993">http://en.wikipedia.org/wiki/Canadian_federal_election,_1993</a>	Politics and government
471	<a href="http://en.wikipedia.org/wiki/Richard_Cordray">http://en.wikipedia.org/wiki/Richard_Cordray</a>	Politics and government
472	<a href="http://en.wikipedia.org/wiki/Don_Dunstan">http://en.wikipedia.org/wiki/Don_Dunstan</a>	Politics and government
473	<a href="http://en.wikipedia.org/wiki/Early_life_and_military_career_of_John_McCain">http://en.wikipedia.org/wiki/Early_life_and_military_career_of_John_McCain</a>	Politics and government
474	<a href="http://en.wikipedia.org/wiki/European_Commission">http://en.wikipedia.org/wiki/European_Commission</a>	Politics and government
475	<a href="http://en.wikipedia.org/wiki/European_Parliament">http://en.wikipedia.org/wiki/European_Parliament</a>	Politics and government
476	<a href="http://en.wikipedia.org/wiki/Fourth_International">http://en.wikipedia.org/wiki/Fourth_International</a>	Politics and government
477	<a href="http://en.wikipedia.org/wiki/Gerald_Ford">http://en.wikipedia.org/wiki/Gerald_Ford</a>	Politics and government
478	<a href="http://en.wikipedia.org/wiki/William_Goebel">http://en.wikipedia.org/wiki/William_Goebel</a>	Politics and government
479	<a href="http://en.wikipedia.org/wiki/Emma_Goldman">http://en.wikipedia.org/wiki/Emma_Goldman</a>	Politics and government
480	<a href="http://en.wikipedia.org/wiki/Herbert_Greenfield">http://en.wikipedia.org/wiki/Herbert_Greenfield</a>	Politics and government
481	<a href="http://en.wikipedia.org/wiki/Benjamin_Harrison">http://en.wikipedia.org/wiki/Benjamin_Harrison</a>	Politics and government
482	<a href="http://en.wikipedia.org/wiki/William_Henry_Harrison">http://en.wikipedia.org/wiki/William_Henry_Harrison</a>	Politics and government

483	<a href="http://en.wikipedia.org/wiki/John_L._Helm">http://en.wikipedia.org/wiki/John_L._Helm</a>	Politics and government
484	<a href="http://en.wikipedia.org/wiki/Her_Majesty%27s_Most_Honourable_Privy_Council">http://en.wikipedia.org/wiki/Her_Majesty%27s_Most_Honourable_Privy_Council</a>	Politics and government
485	<a href="http://en.wikipedia.org/wiki/George_F._Kennan">http://en.wikipedia.org/wiki/George_F._Kennan</a>	Politics and government
486	<a href="http://en.wikipedia.org/wiki/Franklin_Knight_Lane">http://en.wikipedia.org/wiki/Franklin_Knight_Lane</a>	Politics and government
487	<a href="http://en.wikipedia.org/wiki/Terry_Sanford">http://en.wikipedia.org/wiki/Terry_Sanford</a>	Politics and government
488	<a href="http://en.wikipedia.org/wiki/Scottish_Parliament">http://en.wikipedia.org/wiki/Scottish_Parliament</a>	Politics and government
489	<a href="http://en.wikipedia.org/wiki/Solomon_P._Sharp">http://en.wikipedia.org/wiki/Solomon_P._Sharp</a>	Politics and government
490	<a href="http://en.wikipedia.org/wiki/Isaac_Shelby">http://en.wikipedia.org/wiki/Isaac_Shelby</a>	Politics and government
491	<a href="http://en.wikipedia.org/wiki/Arthur_Sifton">http://en.wikipedia.org/wiki/Arthur_Sifton</a>	Politics and government
492	<a href="http://en.wikipedia.org/wiki/South_Australian_state_election,_2006">http://en.wikipedia.org/wiki/South_Australian_state_election,_2006</a>	Politics and government
493	<a href="http://en.wikipedia.org/wiki/Albert_Speer">http://en.wikipedia.org/wiki/Albert_Speer</a>	Politics and government
494	<a href="http://en.wikipedia.org/wiki/State_of_Vietnam_referendum,_1955">http://en.wikipedia.org/wiki/State_of_Vietnam_referendum,_1955</a>	Politics and government
495	<a href="http://en.wikipedia.org/wiki/Ed_Stelmach">http://en.wikipedia.org/wiki/Ed_Stelmach</a>	Politics and government
496	<a href="http://en.wikipedia.org/wiki/Stephen_Colbert_at_the_2006_White_House_Correspondents%27_Association_Dinner">http://en.wikipedia.org/wiki/Stephen_Colbert_at_the_2006_White_House_Correspondents%27_Association_Dinner</a>	Politics and government
497	<a href="http://en.wikipedia.org/wiki/United_Nations_Parliamentary_Assembly">http://en.wikipedia.org/wiki/United_Nations_Parliamentary_Assembly</a>	Politics and government
498	<a href="http://en.wikipedia.org/wiki/Voting_system">http://en.wikipedia.org/wiki/Voting_system</a>	Politics and government
499	<a href="http://en.wikipedia.org/wiki/Rudolf_Wolters">http://en.wikipedia.org/wiki/Rudolf_Wolters</a>	Politics and government
500	<a href="http://en.wikipedia.org/wiki/Alexander_Cameron_Rutherford">http://en.wikipedia.org/wiki/Alexander_Cameron_Rutherford</a>	Politics and government
501	<a href="http://en.wikipedia.org/wiki/1896_Summer_Olympics">http://en.wikipedia.org/wiki/1896_Summer_Olympics</a>	Sport and recreation
502	<a href="http://en.wikipedia.org/wiki/1923_FA_Cup_Final">http://en.wikipedia.org/wiki/1923_FA_Cup_Final</a>	Sport and recreation
503	<a href="http://en.wikipedia.org/wiki/1926_World_Series">http://en.wikipedia.org/wiki/1926_World_Series</a>	Sport and recreation
504	<a href="http://en.wikipedia.org/wiki/1956_FA_Cup_Final">http://en.wikipedia.org/wiki/1956_FA_Cup_Final</a>	Sport and recreation
505	<a href="http://en.wikipedia.org/wiki/1994_San_Marino_Grand_Prix">http://en.wikipedia.org/wiki/1994_San_Marino_Grand_Prix</a>	Sport and recreation
506	<a href="http://en.wikipedia.org/wiki/1995_Japanese_Grand_Prix">http://en.wikipedia.org/wiki/1995_Japanese_Grand_Prix</a>	Sport and recreation
507	<a href="http://en.wikipedia.org/wiki/1995_Pacific_Grand_Prix">http://en.wikipedia.org/wiki/1995_Pacific_Grand_Prix</a>	Sport and recreation
508	<a href="http://en.wikipedia.org/wiki/2000_Sugar_Bowl">http://en.wikipedia.org/wiki/2000_Sugar_Bowl</a>	Sport and recreation
509	<a href="http://en.wikipedia.org/wiki/2003_Insight_Bowl">http://en.wikipedia.org/wiki/2003_Insight_Bowl</a>	Sport and recreation
510	<a href="http://en.wikipedia.org/wiki/2005_ACC_Championship_Game">http://en.wikipedia.org/wiki/2005_ACC_Championship_Game</a>	Sport and recreation
511	<a href="http://en.wikipedia.org/wiki/2005_Sugar_Bowl">http://en.wikipedia.org/wiki/2005_Sugar_Bowl</a>	Sport and recreation
512	<a href="http://en.wikipedia.org/wiki/2005_Texas_Longhorns_football_team">http://en.wikipedia.org/wiki/2005_Texas_Longhorns_football_team</a>	Sport and recreation
513	<a href="http://en.wikipedia.org/wiki/2005_United_States_Grand_Prix">http://en.wikipedia.org/wiki/2005_United_States_Grand_Prix</a>	Sport and recreation
514	<a href="http://en.wikipedia.org/wiki/2006_Chick-fil-A_Bowl">http://en.wikipedia.org/wiki/2006_Chick-fil-A_Bowl</a>	Sport and recreation
515	<a href="http://en.wikipedia.org/wiki/2006_Gator_Bowl">http://en.wikipedia.org/wiki/2006_Gator_Bowl</a>	Sport and recreation
516	<a href="http://en.wikipedia.org/wiki/2007_ACC_Championship_Game">http://en.wikipedia.org/wiki/2007_ACC_Championship_Game</a>	Sport and recreation
517	<a href="http://en.wikipedia.org/wiki/2007_UEFA_Champions_League_Final">http://en.wikipedia.org/wiki/2007_UEFA_Champions_League_Final</a>	Sport and recreation
518	<a href="http://en.wikipedia.org/wiki/2007_USC_Trojans_football_team">http://en.wikipedia.org/wiki/2007_USC_Trojans_football_team</a>	Sport and recreation
519	<a href="http://en.wikipedia.org/wiki/2008_ACC_Championship_Game">http://en.wikipedia.org/wiki/2008_ACC_Championship_Game</a>	Sport and recreation
520	<a href="http://en.wikipedia.org/wiki/2008_Brazilian_Grand_Prix">http://en.wikipedia.org/wiki/2008_Brazilian_Grand_Prix</a>	Sport and recreation
521	<a href="http://en.wikipedia.org/wiki/2008_Humanitarian_Bowl">http://en.wikipedia.org/wiki/2008_Humanitarian_Bowl</a>	Sport and recreation
522	<a href="http://en.wikipedia.org/wiki/2008_Japanese_Grand_Prix">http://en.wikipedia.org/wiki/2008_Japanese_Grand_Prix</a>	Sport and recreation
523	<a href="http://en.wikipedia.org/wiki/2008_Orange_Bowl">http://en.wikipedia.org/wiki/2008_Orange_Bowl</a>	Sport and recreation
524	<a href="http://en.wikipedia.org/wiki/Bids_for_the_2012_Summer_Olympics">http://en.wikipedia.org/wiki/Bids_for_the_2012_Summer_Olympics</a>	Sport and recreation
525	<a href="http://en.wikipedia.org/wiki/Aikido">http://en.wikipedia.org/wiki/Aikido</a>	Sport and recreation
526	<a href="http://en.wikipedia.org/wiki/Amateur_radio_direction_finding">http://en.wikipedia.org/wiki/Amateur_radio_direction_finding</a>	Sport and recreation

527	<a href="http://en.wikipedia.org/wiki/Amateur_radio_in_India">http://en.wikipedia.org/wiki/Amateur_radio_in_India</a>	Sport and recreation
528	<a href="http://en.wikipedia.org/wiki/Arsenal_F.C.">http://en.wikipedia.org/wiki/Arsenal_F.C.</a>	Sport and recreation
529	<a href="http://en.wikipedia.org/wiki/Association_football">http://en.wikipedia.org/wiki/Association_football</a>	Sport and recreation
530	<a href="http://en.wikipedia.org/wiki/Aston_Villa_F.C.">http://en.wikipedia.org/wiki/Aston_Villa_F.C.</a>	Sport and recreation
531	<a href="http://en.wikipedia.org/wiki/Australia_at_the_Winter_Olympics">http://en.wikipedia.org/wiki/Australia_at_the_Winter_Olympics</a>	Sport and recreation
532	<a href="http://en.wikipedia.org/wiki/Sid_Barnes">http://en.wikipedia.org/wiki/Sid_Barnes</a>	Sport and recreation
533	<a href="http://en.wikipedia.org/wiki/Shelton_Benjamin">http://en.wikipedia.org/wiki/Shelton_Benjamin</a>	Sport and recreation
534	<a href="http://en.wikipedia.org/wiki/Moe_Berg">http://en.wikipedia.org/wiki/Moe_Berg</a>	Sport and recreation
535	<a href="http://en.wikipedia.org/wiki/Bodyline">http://en.wikipedia.org/wiki/Bodyline</a>	Sport and recreation
536	<a href="http://en.wikipedia.org/wiki/Luc_Bourdon">http://en.wikipedia.org/wiki/Luc_Bourdon</a>	Sport and recreation
537	<a href="http://en.wikipedia.org/wiki/Brabham">http://en.wikipedia.org/wiki/Brabham</a>	Sport and recreation
538	<a href="http://en.wikipedia.org/wiki/Brabham_BT19">http://en.wikipedia.org/wiki/Brabham_BT19</a>	Sport and recreation
539	<a href="http://en.wikipedia.org/wiki/Donald_Bradman">http://en.wikipedia.org/wiki/Donald_Bradman</a>	Sport and recreation
540	<a href="http://en.wikipedia.org/wiki/Donald_Bradman_with_the_Australian_cricket_team_in_England_in_1948">http://en.wikipedia.org/wiki/Donald_Bradman_with_the_Australian_cricket_team_in_England_in_1948</a>	Sport and recreation
541	<a href="http://en.wikipedia.org/wiki/Eric_Brewer_(ice_hockey)">http://en.wikipedia.org/wiki/Eric_Brewer_(ice_hockey)</a>	Sport and recreation
542	<a href="http://en.wikipedia.org/wiki/Martin_Brodeur">http://en.wikipedia.org/wiki/Martin_Brodeur</a>	Sport and recreation
543	<a href="http://en.wikipedia.org/wiki/Bill_Brown_(cricketer)">http://en.wikipedia.org/wiki/Bill_Brown_(cricketer)</a>	Sport and recreation
544	<a href="http://en.wikipedia.org/wiki/Steve_Bruce">http://en.wikipedia.org/wiki/Steve_Bruce</a>	Sport and recreation
545	<a href="http://en.wikipedia.org/wiki/Simon_Byrne">http://en.wikipedia.org/wiki/Simon_Byrne</a>	Sport and recreation
546	<a href="http://en.wikipedia.org/wiki/Calgary_Flames">http://en.wikipedia.org/wiki/Calgary_Flames</a>	Sport and recreation
547	<a href="http://en.wikipedia.org/wiki/Calgary_Hitmen">http://en.wikipedia.org/wiki/Calgary_Hitmen</a>	Sport and recreation
548	<a href="http://en.wikipedia.org/wiki/Chariot_racing">http://en.wikipedia.org/wiki/Chariot_racing</a>	Sport and recreation
549	<a href="http://en.wikipedia.org/wiki/Central_Coast_Mariners_FC">http://en.wikipedia.org/wiki/Central_Coast_Mariners_FC</a>	Sport and recreation
550	<a href="http://en.wikipedia.org/wiki/Ian_Chappell">http://en.wikipedia.org/wiki/Ian_Chappell</a>	Sport and recreation
551	<a href="http://en.wikipedia.org/wiki/Chelsea_F.C.">http://en.wikipedia.org/wiki/Chelsea_F.C.</a>	Sport and recreation
552	<a href="http://en.wikipedia.org/wiki/Chess">http://en.wikipedia.org/wiki/Chess</a>	Sport and recreation
553	<a href="http://en.wikipedia.org/wiki/Chicago_Bears">http://en.wikipedia.org/wiki/Chicago_Bears</a>	Sport and recreation
554	<a href="http://en.wikipedia.org/wiki/City_of_Manchester_Stadium">http://en.wikipedia.org/wiki/City_of_Manchester_Stadium</a>	Sport and recreation
555	<a href="http://en.wikipedia.org/wiki/Paul_Collingwood">http://en.wikipedia.org/wiki/Paul_Collingwood</a>	Sport and recreation
556	<a href="http://en.wikipedia.org/wiki/A._E._J._Collins">http://en.wikipedia.org/wiki/A._E._J._Collins</a>	Sport and recreation
557	<a href="http://en.wikipedia.org/wiki/Ian_Craig">http://en.wikipedia.org/wiki/Ian_Craig</a>	Sport and recreation
558	<a href="http://en.wikipedia.org/wiki/Cricket_World_Cup">http://en.wikipedia.org/wiki/Cricket_World_Cup</a>	Sport and recreation
559	<a href="http://en.wikipedia.org/wiki/Crusaders_(rugby)">http://en.wikipedia.org/wiki/Crusaders_(rugby)</a>	Sport and recreation
560	<a href="http://en.wikipedia.org/wiki/Cycling_at_the_2008_Summer_Olympics_%E2%80%93_Men%27s_road_race">http://en.wikipedia.org/wiki/Cycling_at_the_2008_Summer_Olympics_%E2%80%93_Men%27s_road_race</a>	Sport and recreation
561	<a href="http://en.wikipedia.org/wiki/December_to_Dismember_(2006)">http://en.wikipedia.org/wiki/December_to_Dismember_(2006)</a>	Sport and recreation
562	<a href="http://en.wikipedia.org/wiki/Derry_City_F.C.">http://en.wikipedia.org/wiki/Derry_City_F.C.</a>	Sport and recreation
563	<a href="http://en.wikipedia.org/wiki/Dover_Athletic_F.C.">http://en.wikipedia.org/wiki/Dover_Athletic_F.C.</a>	Sport and recreation
564	<a href="http://en.wikipedia.org/wiki/Tim_Duncan">http://en.wikipedia.org/wiki/Tim_Duncan</a>	Sport and recreation
565	<a href="http://en.wikipedia.org/wiki/Dungeons_%26_Dragons">http://en.wikipedia.org/wiki/Dungeons_%26_Dragons</a>	Sport and recreation
566	<a href="http://en.wikipedia.org/wiki/Dr_Pepper_Ballpark">http://en.wikipedia.org/wiki/Dr_Pepper_Ballpark</a>	Sport and recreation
567	<a href="http://en.wikipedia.org/wiki/Easy_Jet">http://en.wikipedia.org/wiki/Easy_Jet</a>	Sport and recreation
568	<a href="http://en.wikipedia.org/wiki/Bobby_Eaton">http://en.wikipedia.org/wiki/Bobby_Eaton</a>	Sport and recreation
569	<a href="http://en.wikipedia.org/wiki/Duncan_Edwards">http://en.wikipedia.org/wiki/Duncan_Edwards</a>	Sport and recreation
570	<a href="http://en.wikipedia.org/wiki/Ray_Emery">http://en.wikipedia.org/wiki/Ray_Emery</a>	Sport and recreation

571	<a href="http://en.wikipedia.org/wiki/England_national_football_team_manager">http://en.wikipedia.org/wiki/England_national_football_team_manager</a>	Sport and recreation
572	<a href="http://en.wikipedia.org/wiki/England_national_rugby_union_team">http://en.wikipedia.org/wiki/England_national_rugby_union_team</a>	Sport and recreation
573	<a href="http://en.wikipedia.org/wiki/Everton_F.C.">http://en.wikipedia.org/wiki/Everton_F.C.</a>	Sport and recreation
574	<a href="http://en.wikipedia.org/wiki/FIFA_World_Cup">http://en.wikipedia.org/wiki/FIFA_World_Cup</a>	Sport and recreation
575	<a href="http://en.wikipedia.org/wiki/Fighting_in_ice_hockey">http://en.wikipedia.org/wiki/Fighting_in_ice_hockey</a>	Sport and recreation
576	<a href="http://en.wikipedia.org/wiki/First-move_advantage_in_chess">http://en.wikipedia.org/wiki/First-move_advantage_in_chess</a>	Sport and recreation
577	<a href="http://en.wikipedia.org/wiki/France_national_rugby_union_team">http://en.wikipedia.org/wiki/France_national_rugby_union_team</a>	Sport and recreation
578	<a href="http://en.wikipedia.org/wiki/German_women%27s_national_football_team">http://en.wikipedia.org/wiki/German_women%27s_national_football_team</a>	Sport and recreation
579	<a href="http://en.wikipedia.org/wiki/Adam_Gilchrist">http://en.wikipedia.org/wiki/Adam_Gilchrist</a>	Sport and recreation
580	<a href="http://en.wikipedia.org/wiki/Gillingham_F.C.">http://en.wikipedia.org/wiki/Gillingham_F.C.</a>	Sport and recreation
581	<a href="http://en.wikipedia.org/wiki/Gliding">http://en.wikipedia.org/wiki/Gliding</a>	Sport and recreation
582	<a href="http://en.wikipedia.org/wiki/Go_Man_Go">http://en.wikipedia.org/wiki/Go_Man_Go</a>	Sport and recreation
583	<a href="http://en.wikipedia.org/wiki/Michael_Gomez">http://en.wikipedia.org/wiki/Michael_Gomez</a>	Sport and recreation
584	<a href="http://en.wikipedia.org/wiki/George_H._D._Gossip">http://en.wikipedia.org/wiki/George_H._D._Gossip</a>	Sport and recreation
585	<a href="http://en.wikipedia.org/wiki/The_Great_American_Bash_(2005)">http://en.wikipedia.org/wiki/The_Great_American_Bash_(2005)</a>	Sport and recreation
586	<a href="http://en.wikipedia.org/wiki/Wayne_Gretzky">http://en.wikipedia.org/wiki/Wayne_Gretzky</a>	Sport and recreation
587	<a href="http://en.wikipedia.org/wiki/Orval_Grove">http://en.wikipedia.org/wiki/Orval_Grove</a>	Sport and recreation
588	<a href="http://en.wikipedia.org/wiki/Hare_coursing">http://en.wikipedia.org/wiki/Hare_coursing</a>	Sport and recreation
589	<a href="http://en.wikipedia.org/wiki/Dominik_Ha%C5%A1ek">http://en.wikipedia.org/wiki/Dominik_Ha%C5%A1ek</a>	Sport and recreation
590	<a href="http://en.wikipedia.org/wiki/Thierry_Henry">http://en.wikipedia.org/wiki/Thierry_Henry</a>	Sport and recreation
591	<a href="http://en.wikipedia.org/wiki/Clem_Hill">http://en.wikipedia.org/wiki/Clem_Hill</a>	Sport and recreation
592	<a href="http://en.wikipedia.org/wiki/Damon_Hill">http://en.wikipedia.org/wiki/Damon_Hill</a>	Sport and recreation
593	<a href="http://en.wikipedia.org/wiki/History_of_American_football">http://en.wikipedia.org/wiki/History_of_American_football</a>	Sport and recreation
594	<a href="http://en.wikipedia.org/wiki/History_of_Arsenal_F.C._(1886%E2%80%931966)">http://en.wikipedia.org/wiki/History_of_Arsenal_F.C._(1886%E2%80%931966)</a>	Sport and recreation
595	<a href="http://en.wikipedia.org/wiki/History_of_Aston_Villa_F.C._(1961%E2%80%93present)">http://en.wikipedia.org/wiki/History_of_Aston_Villa_F.C._(1961%E2%80%93present)</a>	Sport and recreation
596	<a href="http://en.wikipedia.org/wiki/History_of_Bradford_City_A.F.C.">http://en.wikipedia.org/wiki/History_of_Bradford_City_A.F.C.</a>	Sport and recreation
597	<a href="http://en.wikipedia.org/wiki/History_of_Gillingham_F.C.">http://en.wikipedia.org/wiki/History_of_Gillingham_F.C.</a>	Sport and recreation
598	<a href="http://en.wikipedia.org/wiki/History_of_Ipswich_Town_F.C.">http://en.wikipedia.org/wiki/History_of_Ipswich_Town_F.C.</a>	Sport and recreation
599	<a href="http://en.wikipedia.org/wiki/History_of_the_National_Hockey_League_(1917%E2%80%931942)">http://en.wikipedia.org/wiki/History_of_the_National_Hockey_League_(1917%E2%80%931942)</a>	Sport and recreation
600	<a href="http://en.wikipedia.org/wiki/History_of_the_National_Hockey_League_(1942%E2%80%931967)">http://en.wikipedia.org/wiki/History_of_the_National_Hockey_League_(1942%E2%80%931967)</a>	Sport and recreation
601	<a href="http://en.wikipedia.org/wiki/History_of_the_National_Hockey_League_(1967%E2%80%931992)">http://en.wikipedia.org/wiki/History_of_the_National_Hockey_League_(1967%E2%80%931992)</a>	Sport and recreation
602	<a href="http://en.wikipedia.org/wiki/Hockey_Hall_of_Fame">http://en.wikipedia.org/wiki/Hockey_Hall_of_Fame</a>	Sport and recreation
603	<a href="http://en.wikipedia.org/wiki/Art_Houtteman">http://en.wikipedia.org/wiki/Art_Houtteman</a>	Sport and recreation
604	<a href="http://en.wikipedia.org/wiki/Karmichael_Hunt">http://en.wikipedia.org/wiki/Karmichael_Hunt</a>	Sport and recreation
605	<a href="http://en.wikipedia.org/wiki/Archie_Jackson">http://en.wikipedia.org/wiki/Archie_Jackson</a>	Sport and recreation
606	<a href="http://en.wikipedia.org/wiki/Jesus_College_Boat_Club_(Oxford)">http://en.wikipedia.org/wiki/Jesus_College_Boat_Club_(Oxford)</a>	Sport and recreation
607	<a href="http://en.wikipedia.org/wiki/Ian_Johnson_(cricketer)">http://en.wikipedia.org/wiki/Ian_Johnson_(cricketer)</a>	Sport and recreation
608	<a href="http://en.wikipedia.org/wiki/Magic_Johnson">http://en.wikipedia.org/wiki/Magic_Johnson</a>	Sport and recreation
609	<a href="http://en.wikipedia.org/wiki/Michael_Jordan">http://en.wikipedia.org/wiki/Michael_Jordan</a>	Sport and recreation
610	<a href="http://en.wikipedia.org/wiki/SummerSlam_(2003)">http://en.wikipedia.org/wiki/SummerSlam_(2003)</a>	Sport and recreation

## APPENDIX D

### THE PENN TREEBANK ENGLISH POS TAG SET AND THEIR MAPPINGS

No	POS	Tag	Mapped to
1	CC	Coordinating conjunction	
2	CD	Cardinal number	
3	DT	Determiner	
4	EX	Existential there	
5	FW	Foreign word	Noun
6	IN	Preposition or subordinating conjunction	
7	JJ	Adjective	Adjective
8	JJR	Adjective, comparative	Adjective
9	JJS	Adjective, superlative	Adjective
10	LS	List item marker	
11	MD	Modal	
12	NN	Noun, singular or mass	Noun
13	NNS	Noun, plural	Noun
14	NNP	Proper noun, singular	Noun
15	NNPS	Proper noun, plural	Noun
16	PDT	Predeterminer	
17	POS	Possessive ending	
18	PRP	Personal pronoun	
19	PRP\$	Possessive pronoun	
20	RB	Adverb	Adverb
21	RBR	Adverb, comparative	Adverb
22	RBS	Adverb, superlative	Adverb
23	RP	Particle	
24	SYM	Symbol	
25	TO	to	
26	UH	Interjection	
27	VB	Verb, base form	Verb
28	VBD	Verb, past tense	Verb
29	VBG	Verb, gerund or present participle	Verb
30	VBN	Verb, past participle	Verb
31	VBP	Verb, non-3rd person singular present	Verb
32	VBZ	Verb, 3rd person singular present	Verb
33	WDT	Wh-determiner	
34	WP	Wh-pronoun	
35	WP\$	Possessive wh-pronoun	
36	WRB	Wh-adverb	

## APPENDIX E

**EXAMPLES OF ORIGINAL/PLAGIARIZED SENTENCE PAIRS AND THE  
CORRESPONDING SIMILARITIES BASED ON EQUATION 3.6.**

NO	Original sentence	Plagiarized sentence	Sim
1	These factors include the condition of a bridge, age, size and complexity; traffic density; impacts of traffic disruption; availability of personnel and equipment; environmental conditions; geographic location; and, construction methods.	These factors include the bridge status, complexness, volume, construction period, and denseness;; impacts of traffic disturbance, availability of workers and equipment ,environmental circumstances; geographic localization.	0.9721
2	There are a lot of technical challenges in designing MANETs, and for a lot of those challenges, solutions have been presented.	In designing MANETs many technical challenges exist, and many have been solved.	0.8469
3	This makes it hard to reiterate research and to fully infer and correctly represent their results.	This makes it difficult to repeat experiments and to fully understand and correctly interpret their results.	0.9643
4	This traditional search method often results in sub-optimal solutions due to inherent limitations in incomplete knowledge representation and the fact that elaborate exploration of the design space is inhibited.	Given the fact that detailed exploration of the search space is restrained and due to underlying restrictions in insufficient knowledge representation, this conventional search method often results in incomplete solutions.	0.8114
5	A two level problem is described here where the subsystem is considered as the low-level problem and the system level (which acts as the coordinator) is considered as the high-level problem.	When describing a two level problem the subordinate denotes the subsystem problem and the system level (this is known as the coordinator) is the upper-level one.	0.8321
6	<i>Engineers, designers</i> and in general, <i>practitioners</i> all influence the knowledge that is brought to solve complex real-life problems.	<i>Professionals</i> from different backgrounds all influence the knowledge that is brought to solve complex real-life problems.	0.9087
7	The most essential point is how to realize an artifactual system that achieves its purpose in unpredictable conditions.	It is important that the artifactual system that accomplishes its purpose in indeterminable situations have to be realized.	0.9606
8	Synthesis is a necessary component of problem solving processes in almost all phases of artifact lifecycle, starting from design, planning, production and consumption until the disposal of the product.	Synthesis is a necessary component of problem solving processes in almost all phases of artifact lifecycle.	0.9218
9	On the other hand, synthesis is described as putting together of parts or elements so as to form a whole, or the combination of separate elements of thought into a whole, as of simple into complex conceptions, species into genera, individual propositions into systems.	Synthesis is defined as the combination of separate and simple elements into a whole, species into genera, and so on.	0.8236

10	Now, the central question is how one can solve the problem of synthesis.	Now, the central question is how one can solve the problem of synthesis: how to determine the system's structure in order to realize its function to achieve a purpose under the constraints of a certain environment.	0.8083
11	It is also argued that, analysis and synthesis, though commonly treated as two different methods, are, if properly understood, only the two necessary parts of the same method.	Synthesis and analysis are two necessary parts of the same method and should not be treated as different.	0.8239
12	The usage of the term 'synthesis' here is somewhat different from the above description, although it is not contradictory to it.	The usage of the term 'synthesis' here is similar to the above description.	0.7819
13	The synthesis is more clearly related to human activities for creation of artificial things, while analysis is related to understanding natural things.	Synthesis is the human activity of artificial creation of things while analysis is related to the natural understanding.	0.9679
14	Analysis is an <i>effective</i> method to clarify the causality of existing natural systems in such fields like physics.	Analysis is a <i>bad</i> technique to explain the relation of existing natural systems in such fields like chemistry.	0.8889
14	Analysis is an effective method to clarify the causality of existing natural systems in such fields like physics.	Analysis is a <i>good</i> technique to explain the relation of existing natural systems in such fields like chemistry.	0.9308
14	Analysis is an effective method to clarify the causality of existing natural systems in such fields like physics.	Analysis is an <i>efficient</i> technique to explain the relation of existing natural systems in such fields like chemistry.	0.9325
14	Analysis is an effective method to clarify the causality of existing natural systems in such fields like <i>physics</i> .	Analysis is an efficient technique to explain the relation of existing natural systems in such fields like <i>biomedicine</i> .	0.8896
14	Analysis is an effective method to clarify the causality of existing natural systems in such fields like physics.	Analysis is an efficient technique to explain the relation of existing natural systems in such fields like <i>medicine</i> .	0.9370
14	Analysis is an effective method to clarify the causality of existing natural systems in such fields like physics.	Analysis is an efficient technique to explain the relation of existing natural systems in such fields like <i>astronomy</i> .	0.9594
14	Analysis is an effective method to <i>clarify</i> the causality of existing natural systems in such fields like physics.	Analysis is an efficient technique to <i>present</i> the relation of existing natural systems in such fields like astronomy.	0.9239
14	Analysis is an effective method to clarify the causality of existing natural systems in such fields like physics.	Analysis is an efficient technique to <i>prove</i> the relation of existing natural systems in such fields like astronomy.	0.9324
14	Analysis is an effective method to clarify the causality of existing natural systems in such fields like physics.	Analysis is an efficient technique to <i>justify</i> the relation of existing natural systems in such fields like astronomy.	0.8791
14	Analysis is an effective method to clarify the causality of existing natural systems in such fields like physics.	Analysis is an efficient technique to <i>disprove</i> the relation of existing natural systems in such fields like astronomy.	0.7463
15	They permit the users to structure the use of the information according to their specific social relationships.	They allow the users to organize the use of the information according to their particular cultural interest.	0.9408
16	Successful research efforts not only impact information management tasks, but can be extended to support knowledge discovery and dissemination.	Productive research endeavors affect information management and can be extended to handle knowledge discovery and dissemination.	0.9699

17	Examples of such research include the application of INQUERY to support the full-text search of environmental regulations or the use of arbitrarily structured metadata to mark documents for data search and exchange.	For instance the INQUERY project allows full-text search of environmental rules or the use of randomly structured information to differentiate documents.	0.9579
18	The second type of text analysis research in AEC attempts to develop domain-specific linguistic resources by analyzing text corpora in support of general information management tasks.	The second kind of AEC research examines text collections in an effort to build up domain-specific linguistic tools with respect to all-purpose tasks.	0.9124
19	Such research might use controlled vocabularies to integrate heterogeneous data representations into product models and , or automatically suggest keywords for construction procurement applications.	Such research is utilized by construction acquisition applications by automatically indicates keywords from specific knowledge that combine non-uniform representations.	0.9099
20	The third kind of research in AEC suggests several schemes to construct the membership functions between desired information requests and sources for IR applications.	The third type of research in AEC proposes various strategies to build the mapping functions between information requests and desired information sources for IR applications.	0.9683
21	A <i>larger</i> proportion of past research is of this type.	A <i>larger</i> amount of previous research is of this kind	0.9607
21	A larger proportion of past research is of this type.	A <i>large</i> amount of previous research is of this kind.	0.9580
21	A larger proportion of past research is of this type.	A <i>huge</i> amount of previous research is of this kind.	0.7853
22	This suggests that the scale of reference collections for AEC applications might not be as critical as it is in general information science research, as long as it addresses the characteristics of the targeted information sources.	This indicates that the size of source corpora for AEC systems might not be as important as it is in all-purpose information science research, as long as it covers the features of the aimed information sources.	0.9521
23	Past research shows a trend to developing AEC-specific semantic/linguistic resources that are specially designed to support the operations of text retrieval.	Previous research reveals a tendency to emerging AEC-specific semantic/linguistic resources that are particularly intended to hold the processes of text retrieval.	0.9481
24	A significant amount of domain information is located in text documents, images, audio and video recordings, and project schedules, all of which may exist outside of the traditional database model.	Database model is not the only source of information, a large amount of data resides in text document, images, audio and video recording, and project schedules.	0.8638
25	Because of those complex data structures, researchers are increasingly adopting IT to cope with these non-structured data formats.	The non-structured data formats have lead researchers to adopt IT to handle this problem.	0.9337
26	McKechnie et al. applied machine learning methods to aid human bibliographers in <i>classifying</i> documents.	McKechnie et al. employed machine learning approaches to assist bibliographers in <i>classifying</i> documents.	0.9423
26	McKechnie et al. applied machine learning methods to aid human bibliographers in classifying documents.	McKechnie et al. employed machine learning approaches to assist bibliographers in <i>categorizing</i> documents.	0.9342

26	McKechnie et al. applied machine learning methods to aid human bibliographers in classifying documents.	McKechnie et al. employed machine learning approaches to assist bibliographers in <i>summarizing</i> documents.	0.8081
27	<i>Researchers have attempted to apply agent technology to manufacturing enterprise integration, enterprise collaboration (including supply chain management and virtual enterprises), manufacturing process planning and scheduling, shop floor control, and to holonic manufacturing as an implementation methodology.</i>	Researchers have attempted to apply agent technology to manufacturing enterprise <i>collaboration</i> .	0.7980
27	Researchers have attempted to apply agent technology to manufacturing enterprise integration, enterprise collaboration (including supply chain management and virtual enterprises), manufacturing process planning and scheduling, shop floor control, and to holonic manufacturing as an implementation methodology.	Researchers have attempted to apply agent technology to manufacturing enterprise <i>activities</i> .	0.7907
27	Researchers have attempted to apply agent technology to manufacturing enterprise integration, enterprise collaboration (including supply chain management and virtual enterprises), manufacturing process planning and scheduling, shop floor control, and to holonic manufacturing as an implementation methodology.	Researchers have attempted to apply agent technology to manufacturing enterprise <i>cooperation</i> .	0.8125
28	Detection of foldable subunits in proteins is an important approach to understand their evolutions and find building motifs for de novo protein design.	Proteins' foldable subunits identification is crucial for recognizing the primitive themes for de novo protein structure and evolution.	0.9388
29	In the supply chain library proposed by Swaminathan et al. , two categories of elements are distinguished: structural elements and control elements, where structural elements refer to the production entities (retailers, distribution centers, plants, suppliers, transportations) and control elements are those helping in coordinating flow of products by efficient message interactions (inventory, demand, supply, flow and information controls).	In the supply chain library suggested by Swaminathan et al. , two types of elements are identified: control and structural elements, where control elements are those aiding in managing flow of products by effective message communications (information controls , flow , demand, inventory, and supply) and structural elements denotes to the production entities (transportations , distribution, suppliers, plants ,retailers, and centers).	0.9418
29	This suggests that this protein may be divided into two foldable halves.	This indicates that this protein may be separated into two equal foldable fractions.	0.9299
30	Successful research efforts not only impact information management tasks, but can be extended to support knowledge discovery and dissemination.	Managing information is not the net effect of the success of such efforts, knowledge discovering and disseminating and are also consequences of this success.	0.5054

31	The increasing importance of text-based information retrieval (IR) developments in the architecture, engineering, and construction industries (AEC) and the lack of sharable testing resources to support these developments call for an approach that can be used to generate domain-specific reference collections.	Information retrieval (IR) developments are important in many fields.	0.6476
32	Past AEC text collections shows that most of the listed research did not attempt to use a testing environment that mimics the web, an enormous document space even if some documents were originally web pages.	An environment that simulates the Web has not been used in previous AEC research.	0.8265
33	These practices create use cases for the text-based IR applications in AEC, which mainly target information systems whose collection sizes are limited.	This had led to AEC models intended for systems with small domain properties.	0.7872
34	This suggests that the scale of reference collections for AEC applications might not be as critical as it is in general information science research, as long as it addresses the characteristics of the targeted information sources.	The corpus size is not much important in AEC applications.	0.7793
35	A second observation of this past research reveals that several research efforts were dedicated to creating and utilizing linguistic resources such as keywords or synonyms in order to support query formulation or search evaluation.	Semantics have been utilized extensively in previous research to support query formulation.	0.6077
36	In addition, domain concepts organized in the form of a taxonomy, thesaurus, or ontology were heavily applied, as evidenced by the many past research efforts that have built their search methodologies upon classification systems.	Ontology has been applied extensively in previous research to support system classification.	0.6391
37	Past research shows a trend to developing AEC-specific semantic/linguistic resources that are specially designed to support the operations of text retrieval.	AEC-specific resources that support information retrieval were the focus of previous research.	0.6363
38	There are a lot of technical challenges in designing MANETs, and for a lot of those challenges, solutions have been presented.	Designing MANETs is an area of extensive research.	0.7579
39	However, it has become apparent that simulation can only be a first step in the evaluation of algorithms and protocols for MANETs.	Simulation is one of several processes in designing MANETs.	0.7103
40	Furthermore, users transfer their social behavior increasingly to networks and networked applications.	Moreover users can share their interest over the Internet.	0.8619

41	Automatic resilience, fault management and overload mechanisms have been proposed at different layers: fast reroute mechanisms at the network layer or dependable overlay services for supporting vertical handovers in mobile networks and at the application layer.	Flexibility, error handling, fast reroute mechanisms are separated over the application and network layers.	0.7414
42	MAC layer emulators simply determine the nodes that should receive a given packet: if a node is emulated to be within radio range of another node, a filter tool allows the exchange of packets between them, if the nodes are out of each others range, the respective packets are dropped.	In MAC layer, node A can receive a packet from another node B if an emulator decided during the filter process that B is within the same frequency of A otherwise it will be excluded.	0.8670
43	The authors ran several experiments with OLSR and AODV.	They had made considerable and significant efforts and various tests in conducting their findings with AODV and OLSR.	0.8373
44	Recently EC-based approaches have been applied to several paper processing problems.	EC-based methods are used lately in various information retrieval tasks.	0.8658
45	It is based on the ideas inspired by biology, like self-organization, evolution, learning and adaptation.	Its processes acquired from biology.	0.7394
46	The maximum communication distance is defined as the point where the packet reception probability drops below 85%.	The connection range ends when the possibility of losing packets becomes less than 85%.	0.8827
47	Due to the different transmission range, the AODV timers had to be adapted.	The AODV must confirm to the changes in communication distance.	0.5579
48	Application of EC techniques to this class of problem is growing, but has found limited application in chemical engineering.	Contrasting to chemical engineering, EC methods are achieving considerable attention in this type of tasks.	0.6585
49	TFIDF vector model uses term frequency (TF) and inverse document frequency (IDF) to measure how important a word is to a document in the collection .	A vector entry in the TFIDF model is the multiplication of term occurrence in a document (TF) and its reciprocal count in the corpus (IDF).	0.8087
50	The Okapi model treats term occurrence as a probability problem and calculates the similarity between queries and documents to generate ranked results.	Okapi is a probabilistic model that measures the relevancy likelihood between query terms and documents.	0.7247
51	The Mediator approach is another type of federation architecture.	The Mediator is a union of various systems.	0.8319