Poach Tracker: An Online Plagiarism Detection Tool

Aaisha Anjum
IMS Engineering College

Avantika Srivastava
IMS Engineering College

Sonal Shalya
IMS Engineering College

Kajal Goel
IMS Engineering College

Avdhesh Gupta, PhD
Associate Professor
IMS Engineering College

ABSTRACT

Plagiarism refers to “the act of copying material without actually conceding the original source”. Plagiarism has seen a wide spread activity in the recent times. The increased in the number of materials available now in the electronic form and the easy access to the internet has increased plagiarism. Various techniques are available which help us to detect plagiarism. This paper proposes algorithm foe plagiarism detection over web using semantic networks. It also shows that a proposed method is in general capable for retrieving the source document from the web using a search engine API when sentences are being infringed. It also calculates the threshold value for different URLs.

Keywords

Plagiarism, Plagiarism Detection, Knuth MorrisPratt, Threshold Value

1. INTRODUCTION

The World Wide Web (WWW) is the biggest source of information these days. Availability of documents has increased in the WWW and the ease to access these documents has lead to a serious problem of using others work without giving credits. The ease of such access and browse web pages to get the information has made today’s life more comfortable, it would be very difficult to imagine the academic research without the internet and web. Now, it is also very easy to use someone else work easily illegally or intelligently without citing credit to the original writer. This is the problem of Plagiarism.

Plagiarism is the act to use someone else’s work and ideas without giving due rights to the original writer and regard the work as your own. It is not sufficient now that leaning only on exact word or phrase matching for plagiarism detection. People declare themselves as authors of the material by paraphrasing or rearranging words to give new look to their sentences.

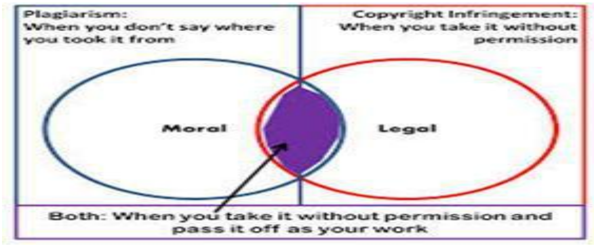

Plagiarism:
When you don't say where you took it from
Copyright Infringement:
When you take it without permission
Moral
Legal
Both: When you take it without permission and pass it off as your work

Fig. 1: Defining plagiarism

Many students make (intentionally or unintentionally) some type of cheating and plagiarism in their assignments which makes difficult for the teachers to detect plagiarism in student’s assignment by hand. The detection process becomes easier, faster and more efficient if it is performed automatically. It is often hard to reveal plagiarism because many methods have been developed to detect some instances of plagiarism such as changing the structure of sentences or when replacing words slightly by synonyms or when the copied sentences are deliberately modified. A proposed method based on extracting name entities and common nouns is in general capable of retrieving the source document from the web using search engine API when sentences are being moderately plagiarized. The process of web based plagiarism detection is first of all we select the target file by pressing the browse button which we want to check for plagiarism. After selecting it will be checked on the web and different links will be provided from where it has been plagiarized and the threshold value for each of the ink will be calculated.

To provide an access to a large number of web documents there are two methods:- First method is by utilizing General purpose search engines (like Google, Yahoo, Bing) etc as they provide access services to their system. The traditional method for measuring the similarity between the document and vulnerable to fail in some complex plagiarism patterns and it is necessary to incorporate semantic based techniques for more accurate plagiarism detection. The suspected document can be considered as a sequence of queries submitted to the search engine; the result can be then compared with the input document. The main idea is to analyze the grammar of target documents and to find irregularities within the syntax of sentences, regardless of the usage of concrete words. If the suspicious sentences are found then the string matching algorithm tries to select and combine those sentences into potentially plagiarized sections.

2. LITERATURE REVIEW

First, In 2006 Maurer et al. [1] classify the plagiarism detection method in to three categories namely Stylometry, Document Comparison and Web Searching. In Stylometry analysis plagiarism is found by the author writing style and this quantification is done by some statistical method without external source known as ‘intrinsic plagiarism detection[2].Every author has its own writing style, if the style of writing is changed with successive sentences then plagiarism is there in the document [3].this technique is not so popular because of no original document is available to support the fact of plagiarism [4].

In document comparison method [4], the plagiarism can be detected by either syntactically or semantically. Semantically plagiarism detection is done by finding the similar words and sentences which was modified by their synonyms [5]. Kang et al. [6] proposed the PPChecker which calculate the data copied from the original text to the query document based on the linguistic pattern namely: the exact sentence copying, word deletion, word substitution, word insertion and whole sentence change pattern. Other approach for semantic analysis was proposed by Tachaphetpiboon et al. [7] in which grammar rule was identify by the use of parsing in the text document and after that these grammar rule was compared to the structure of text. Semantic Sequence Kin (SKK) approach for semantic analysis was proposed by Bao et al. [8] which use the word position information for plagiarism detection. In syntactic plagiarism detection meaning of the word, phrase or sentence is not considered. Shiva Kumar et al. [9] proposed SCAM for plagiarism detection in the sentence which measure the global similarity but can't process the positional information of copied content.

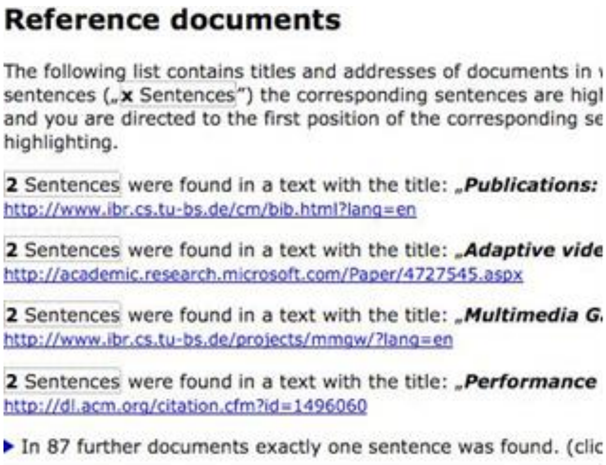
The search engine API is the core of many web-based plagiarism detection techniques. Web based plagiarism detection tool is further categorized in to server side and client side plagiarism detection tool. There are many freely available tools in the market with paid one such as DocCop [10], Plagium [11], Turnitin [12], Safe-assignment [13], Urkund [14] are the tools of server side or based on web servie. Except these CopyCatch [15], WCopyfind [16], EVE 2[17], MOSS [18] and GPSP [19] are the tools available on web with client side functionality.

3. AVAILABLE TOOLS

To avoid the academic dishonesty there are several tools available. Some of them are discussed in this section. Doc Cop [10] is a web based plagiarism and collusion detection tool which breaks the query document in to N-gram phrases and then measure the plagiarism by conducting searching in google for each phrases [4]. It measure the similarity between the document and the web. Plagium [11] is another freely available tool. According to [4], Plagium perform better than Doc Cop. It is also based on search engine API.

Trunitin [12] is the web based commercial product from iParadigms. In which detection and processing is done remotely. Trunitin system database contain approximately 4.5 billion pages of books and journals. Safe- assignment [13] is the web based service provided by the Mydropbox, which covers 8 billion documents from ProQuest, FindArticles& other major scholastic databases [1]. Institute ForAngewandteLerntechnologien (IFALT) gives another web based service known as Docoloc. It also utilizes the searching and ranking of Google API. Urkund [14] is one more plagiarism detection technique based on web service which uses the email system for submitting and viewing the result.

CopyCatch [15] is a client based tool which is used to compare locally available database of documents. WCopyfind [16] is open source tool for detecting words or phrases of defined length within local repository of documents. EVE2 (Essay Verification Engine) [17] is an another client based tool which is based on own internet search mechanism to find the plagiarism data. Except these GPSP [18], MOSS [19] and JPlag are the other tools available on web.



Reference documents

The following list contains titles and addresses of documents in v
sentences („x Sentences") the corresponding sentences are higl
and you are directed to the first position of the corresponding se
highlighting.

2 Sentences were found in a text with the title: ***„Publications:***
http://www.ibr.cs.tu-bs.de/cm/bib.html?lang=en

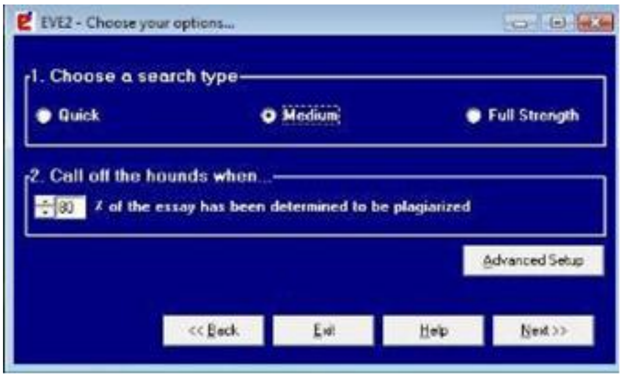
2 Sentences were found in a text with the title: ***„Adaptive vide***
http://academic.research.microsoft.com/Paper/4727545.aspx

2 Sentences were found in a text with the title: ***„Multimedia G***
http://www.ibr.cs.tu-bs.de/projects/mmgw/?lang=en

2 Sentences were found in a text with the title: ***„Performance***
http://dl.acm.org/citation.cfm?id=1496060

▶ In 87 further documents exactly one sentence was found. (clic

Fig. 2: An example of plagiarism report by Docoloc



EVE2 - Choose your options...

1. Choose a search type

○ Quick ● Medium ○ Full Strength

2. Call off the hounds when...

80 % of the essay has been determined to be plagiarized

Fig. 3: the interface of EVE 2 web searching

4. IMPLEMENTATION

Most web based Plagiarism detection tools use search engine APIs. The semantic relatedness approach based on the work will be adopted in measuring the similarity between sentences by adding supports for other parts-of-speeches in particular for adjectives and adverbs.

A. Document Preprocessing

Document Preprocessing involves following stages for all query documents-

Tokens which are non-essential such as numbers parenthesis and punctuations are excluded and the sentences whose length is less than three are not considered automatically.

Stop words are also omitted.

All functional words such as conjunction, preposition, articles, auxiliary verbs, pronouns and cardinal words are also not considered.

Web document Retrieval

The procedure of retrieving the source document from the web includes:- First of all selecting the target file from the browse option and then search it through Google API. Each source documents URL will be provide and recorded and the objective is to determine the top URL from where most of the content has been retrieved.

It is based on the following matrix:-

The number of URLs returned from all queries.

The minimum number of queries required to retrieve the source document.

Match the query with the selected URL with the help of String matching algorithm.

The number of overall utilized queries by a technique.

Count the number of plagiarized words with the total length of the target string.

Accordingly calculate the Threshold value.

There are three techniques that will be evaluated. The first technique takes every n- consecutive words (greater than 3) from the source document as queries. Queries that are totally stop words will be omitted. 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 rank according to their importance (weights). Each word in the query is weighted according to equation. The query weight is the summation of all its individual words weight. 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 formulated in sub queries in decreasing length with the minimum length of two words.

C. Document comparison

Matching of strings consisted of stages step to find one string or more to all cases at a string (in general is called as pattern string) in text. All matching algorithm of string will yield all pattern string found in text.

Knuth-Morris-Pratt Algorithm:

Knuth Morris Pratt's algorithm for string matching was used for the scheme to detect plagiarism. This method was developed by Donald Knuth and Vaughn Pratt who worked together. This algorithm is known for its linear time exact matching. The algorithm compares the text from left to right and is able to shift more than one position. The algorithm is very clever in a sense that it is able to avoid trivial comparisons due to the preprocessing phase. KMP works by preprocessing the pattern to be searched in a document. This process involves finding the largest prefix, P[0..j-1], that matches the largest suffix, P[1..j]. This value represents the number of shifts to be done when a mismatch, match, or partial match occurs. It indicates how much of the last comparison can be reused if the algorithm fails to find a match. It is highly efficient because the number of comparisons for the pattern against the original text is minimized. In all, the algorithm has a complexity of O(n+m) where O(m) is the computation for the prefix function values and O(n) is the comparison of the pattern to the text. KMP is very advantageous because it never has to backtrack on a string which makes it efficient for large text files.

Table 1: KMP algorithm

j	0	1	2	3	4	5
P[j]	A	B	a	c	a	b
F[j]	0	0	1	0	1	2

P[0]- a:0 //It is zero because there is no prefix or suffix

P[1]- ab(a!=b) -> 0

P[2]- aba(a=a),(ab!=ba) ->1

P[3]- abac(a!=c),(ab!=ac),(aba!=bac) ->0

Calculation of threshold value:

After we have retrieved our plagiarized content by comparing the target document with the source document we calculate the Threshold value. The threshold value is calculated by dividing the total number of plagiarized words to the total length of target document. The value obtained is the percentage of content plagiarized in that document. The value can be calculated the threshold value for all URL according to the specific requirement.

5. RESULTS AND CONCLUSION

The query engine is the component that the user most closely contact with.. By providing a query string, the user requests information that the query engine parses. The query engine retrieves matching inverted lists for the terms the user provided. The search Engine provides results to the user by ranking in decreasing estimated relevance and this relevance is estimated using a similarity measure. Each search result consist of the title of the web document i.e its URL and a short fragment that describes the document. The returned URL of query are then compared to the source query from where the query document are originated to check whether a particular query is successful or not. Threshold value is calculated for URL's and the source document from where the query document has mostly been plagiarized.

Plagiarized sentences cannot reveal sufficiently through syntactic information, this comes from the fact that ordering of information is not so important in computing the similarity between plagiarized sentences. The Semantic relatedness between two sentences that is based on the path length of a semantic relation between their words, the modification of the words , information content of words will increase the overall performance has recall gain will out way precision loss.

Web searching technique using a search engine API is exhausted and unnecessary to retrieve the source document, also has many drawbacks including a large lists of documents to be downloaded, a small fraction of hits over misses. However, this can be avoided by extracting rare queries in the given document or by extracting named entities since these are often hard to be plagiarized. The aim of this application is to look for percentage of similarity two the files.

In future, we intend to integrate the different components of the system to build one final web based plagiarism detection system. We will be thoroughly investigating the performance of different similarity measures before incorporating them in the final similarity computation model.

6. REFERENCES

[1] H. Maurer et al., "Plagiarism- A Survey", Journal of universal computer science, vol. 2,no. 8, 2006.

[2] Eissen& Stein, "Intrinsic Plagiarism Detetion", Proc. Of European conf. on Information Retrieval (ECIR-06), pp. 565-569, Springer,2006.

[3] S. Gruner et al. "Tool support for plagiarism detection in text documents", Proceedings of the 2005 ACM Symposium on Applied Computing, pp. 776 – 781, 2005.

[4] Ahmed Jabr Ahmed Muftah, "Document Plagiarism detection algorithm using semantic networks", M.S. Thesis, University Technology,Malaysia, 2009.

[5] P. Clough, "Old and new challenges in automatic plagiarism detection", Plagiarism Advisory Service, vol. 10, Department of Computer Science, University of Sheffield, 2003.

[6] N. Kang and A. Gelbukh, “PPChecker: Plagiarism Pattern Checker in Document Copy Detection”, In: Sojka, P., Kopecek, I., Pala, K. (eds.)

[7] TSD 2006. LNCS, vol. 4188, pp. 661–667. Springer,Heidelberg. 2006.

[8] S. Tachaphetpiboon,N. Facundes, and T. Amornraksa, “Plagiarism Indication by Syntactic-Semantic Analysis”, Proceedings of Asia-Pacific Conference on Communications 2007.

[9] J. Bao et al., “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.

[10] N. Shivakumar and H. Garcia-Molina, “SCAM: a copy detection mechanism for digital documents”, In Proc. International Conference on Theory and Practice of Digital Libraries, Austin, Texas. 1995.

[11] Available: http://www.doccop.com

[12] Available: http://www.plagium.com

[13] iParadigms anti plagiarism product website, http://www.plagiarism.org/, visited: 22 July 2006.

[14] Mydropbox, SafeAssignment Product Brochure, http://www.mydropbox.com/info/SafeAssignment_Standalone.pdf, visited: 22 July 2006.

[15] Urkund website, http://www.urkund.com/ visited: 22 July 2006.

[16] CopyCatch product website, http://www.copycatchgold.com/, visited: 22 July 2006.

[17] WCopyfindwebsite.http://plagiarism.phys.virginia.edu/Wsoftware. html,visited:22 July 2006.

[18] EVE Plagiarism Detection System website, http://www.canexus.com/eve/, visited: 22 July 2006.

[19] Glatt Plagiarism Services website, http://www.plagiarism.com/, visited: 22 July 2006.

[20] MOSS, A System for Detecting Software Plagiarism website, http://www.cs.berkeley.edu/~aiken/moss.html, visited: 22 July 2006