ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

MINING THE WEB FOR PLAGIARIZED WEB CONTENT USING SIMILARITY MEASURE TECHNIQUE

Prof. M. MATHIYALAGAN[1]
[1]Assistant Professor,
[1]Dept of Computer Science,
[1]Park's College, Tirupur – 5.

Abstract

Two important and active areas of current research are data mining and the World Wide Web. A natural combination of the two areas, sometimes referred to as Web mining, has been the focus of several recent research projects and papers. The heterogeneity and the lack of structure that permeates much of the ever expanding information sources on the World Wide Web, such as hypertext documents, makes automated discovery, organization, and management of Web-based information difficult. As with any emerging research area there is no established vocabulary, leading to confusion when comparing research efforts.

Different terms for the same concept or different definitions being attached to the same word are commonplace. The term Web mining has been used in two distinct ways. Web content mining is the process of extracting knowledge and information discovery from sources across the World Wide Web. Mirrored web pages are very common in internet. Sometimes, without the knowledge and permission of the owners of the original web page, someone may duplicate the contents of the web page in their page. Finding such plagiarism in the vast internet is a challenging task. In this research we explore the web mining technology and a plagiarism detection paradigm for web mining.

A working prototype of the proposed system will be developed partially in C and partially on Matlab. The Integration of the C code with the Matlab code will be done using Matlab Mex DLL interface programming. The performance of the system will be evaluated using suitable Metrics.

Keywords: - Web, Content, Plagiarism, Mining, Research.

1. INTRODUCTION

Data mining is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. It has been defined as, the automated analysis of large or complex data sets in order to discover significant patterns or trends that would otherwise go unrecognized.

The goal of data mining is to unearth relationships in data that may provide useful insights. Data mining tools can sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions, performance bottlenecks in a network system and identifying *anomalous data* that could represent data entry keying errors. The ultimate significance of these patterns will be assessed by a domain expert - a marketing manager or network supervisor - so the results must be presented in a way that human experts can understand.

Data mining tools can also automate the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

The core components of data mining technology have been under development for decades, in research areas such as statistics, artificial intelligence, and machine learning. Today, the maturity of these techniques, coupled with high-performance relational database engines and broad data integration efforts, make these technologies practical for current data warehouse environments.

Web based Learning Management systems are now a mandatory resource in any reputable academic institution. Institutions benefit a great deal from the flexibility that these systems provide to the academic and learner community.

Verifying the quality of submissions received by academic is one of the toughest tasks in today's learning environment. Numerous researchers have documented the extent of plagiarism and student cheating over the past 60 years observes that it is so easy to plagiarize using Internet sources, students may plagiarize without recognizing that they are doing so, knowing that plagiarism is ethically wrong. Both a) and b) come under the category of IPR (Intellectual Property Rights) regulations and plagiarism. An institution needs to adhere to the quality audit process that details the auditing mechanism that must be applied by the academics at the institution while dealing with Digital content in e-learning systems and student submissions.

Many educators acknowledge that more students than ever plagiarize material from different sources, especially the Internet observed that the information technology boom has attributed to the widespread practice of plagiarism.

2. WEB MINING & PLAGIARISM

By looking at web mining from an ethical perspective, we shall discover a field of tension, between advantages on the one hand and disadvantages on the other. As ethics is the branch of philosophy concerned with the nature of morals and moral evaluation, an ethical perspective will raise questions like what is right or wrong, what is beneficial or harmful. Ethical research focuses on three types of problems. First, there are situations in which normative principles are clearly disregarded. Then there are ethical problems concerning new issues (types of problems that do not match existing cases) where it is a question of how traditional principles can be applied. The third type of ethical situations deals with the category of normative conflicts. A normative conflict appears whenever there are both good and bad sides to a matter.

The issue of web mining is a normative conflict where good refers to the benefits of web mining and bad refers to its possible harmful implications, in other words the ethical values that are threatened. Values are core beliefs or desires that guide or motivate attitudes and actions, and determine how people behave in certain situations. As ethics is a reflection on morality, ethical values could be described as that which subjects affirm as moral in human behaviour [Xiaohe, 1998]. Thus ethical values have a normative function and are the motive for moral human behaviour. A value can be seen as a global goal. Such a goal needs to be driven by a means, presented by more specific norms. For instance, the value of privacy is driven by norms like respecting someone's private life and not misusing someone's personal data. Norms would be meaningless without values.

Knowledge discovered after mining the web, could pose a threat to people, when for instance personal data is misused. However, it is this same knowledge factor that can imply lots of different advantages, as it is of high value to all sorts of applications concerning planning and control. Kosala, Blockeel and Neven have already described some specific benefits of web mining, like improving the intelligence of search engines. Web mining can also contribute to marketing intelligence by analyzing the web user's on-line behavior and turning this information into marketing knowledge.

There are different ways to mine the web. To structurally analyse the field of tension we need to be able to distinguish between those different forms of web mining. The different ways to mine the web are closely related to the different types of web data. We can distinguish actual data on web pages, web structure data regarding the hyperlink structure within and across web documents, and web log data regarding the users who browsed the web pages.

2.1. About Plagiarism

Plagiarism means copying work and pretending that it as our work. Generally in all academic institutions Plagiarism is not allowed while doing simple class assignments, essays and projects. Any student who plagiarizes the work of others will get reduction in marks or will get no mark at all.

Generally the following actions are considered to be plagiarism:

- copying paragraphs or programs from a textbook;
- Directly handling someone's source code without mentioning it.
- copying another person's work either with or without their knowledge;
- Working together in groups of two or more to produce a single program or essay and then each member of the group submitting a copy of this as their own work.

A fundamental question in information theory and in computer science is how to measure similarity or the amount of shared information between two sequences. Computer systems for source code plagiarism detection have been in existence for over twenty years (Halstead 1977, Ottenstein 1977) and are now routinely used in many academic institutions.

The first systems were based upon attribute counting algorithms, extracting various superficial metrics from source code submissions, for example a count of the use of a particular reserved word, and then flagging those pairs of submissions which were significantly close for tutors to inspect further. Later developments saw the introduction of structure metric systems where each submission is reduced to a series of identifiers, or tokens, representing, for example, a function call or a variable declaration. Pairs of tokenised submissions are then recursively searched for the longest common token sequences and the proportion of the submissions matched used as a similarity metric. Structure metric systems have been shown to be more effective than attribute counting systems (Verco & Wise 1996) and all existing systems use those principles.

The following algorithms have been used for plagiarism detection. The algorithms normally used in plagiarism detection software are string tiling, Karp-Rabin algorithm, Haeckel's algorithm, k-grams, string matching algorithm, the authors describe two algorithms that they have used to test for efficiency in plagiarism detection.

In the authors propose a system that is based on properties of assignments that course instructors use to judge the similarity of two submissions instead of the popular text-based analyses. This system uses neural network techniques to create a feature-based plagiarism detector and to measure the relevance of each feature in the assessment. The system was trained and tested on assignments from an introductory computer science course, and produced results that are comparable to the most popular plagiarism detectors.

Two popular methods by Levenshtein and Damerau defined edit distances that can be used to compare the similarity of two strings of characters with each other. These distances are used in a variety of applications ranging from DNA analysis to plagiarism detection.

The following are some of the work already done in this field. Scherbinin and Butakov used *Levenshtein* distance to compare word *n*-gram and combine adjacent similar grams into sections. In another approach the *Levenshtein* distance and simplified *Smith-Waterman* algorithm were merged as a single algorithm for the identification and quantification of local similarities in plagiarism detection. In [6] the researchers used the *LCS* distance combined with other POS syntactical features to identify similar strings locally and rank documents globally

A commonly-used bottom-up dynamic programming algorithm for computing the Levenshtein distance involves the use of an (n + 1) × (m + 1) matrix, where n and m are the lengths of the two strings. This algorithm is based on the Wagner-Fischer algorithm for edit distance.

The second algorithm, the Smith-Waterman algorithm is a classical method of comparing two strings with a view to identifying highly similar sections within them. It is widely-used in finding good near-matches, or so-called local alignments, within biological sequences. But now, Smith-Waterman algorithm has been used in the text plagiarism detection, and our paper will simplify it.

3. PROPOSED MODEL FOR WEB PLAGIARISM

Internet Digital content is easy to copy and therefore it is assimilated into a learning material without checking for integrity or authorship; mention cases of academics copying from conference and journal papers. Verifying the quality of submissions received by academic is one of the toughest tasks in today's learning environment. Numerous researchers have documented the extent of plagiarism and student cheating over the past 60 years. observes that it is so easy to plagiarize using Internet sources, students may plagiarize without recognizing that they are doing so, knowing that plagiarism is ethically wrong. Both a) and b) come under the category of IPR (Intellectual Property Rights) regulations and plagiarism. An institution needs to adhere to the quality audit process that details the auditing mechanism that must be applied by the academics at the institution while dealing with Digital content in e-learning systems and student submissions. Many educators acknowledge that more students than ever plagiarize material from different sources, especially the Internet observed that the information technology boom has attributed to the widespread practice of plagiarism.

The range of different types of plagiarism has been by mentioned in namely copy and paste plagiarism, word switch plagiarism, but they essentially mean borrowing ideas without crediting the real author. Academic Institutions try to counter plagiarists by establishing strict academic integrity and anti-plagiarism policies. One approach to fighting would be to change the campus culture from focusing plagiarism on "catching cheaters" to promoting academic integrity.

Plagiarism means copying work and pretending that it as our work. Generally in all academic institutions Plagiarism is not allowed while doing simple class assignments, essays and projects. Any student who plagiarizes the work of others will get reduction in marks or will get no mark at all.

Generally the following actions are considered to be plagiarism:

- copying paragraphs or programs from a textbook;
- Directly handling someone's source code without mentioning it.
- copying another person's work either with or without their knowledge;
- Working together in groups of two or more to produce a single program or essay and then each member of the group submitting a copy of this as their own work.

A fundamental question in information theory and in computer science is how to measure similarity or the amount of shared information between two sequences. Computer systems for source code plagiarism detection have been in existence for over twenty years (Halstead 1977, Ottenstein 1977) and are now routinely used in many academic institutions.

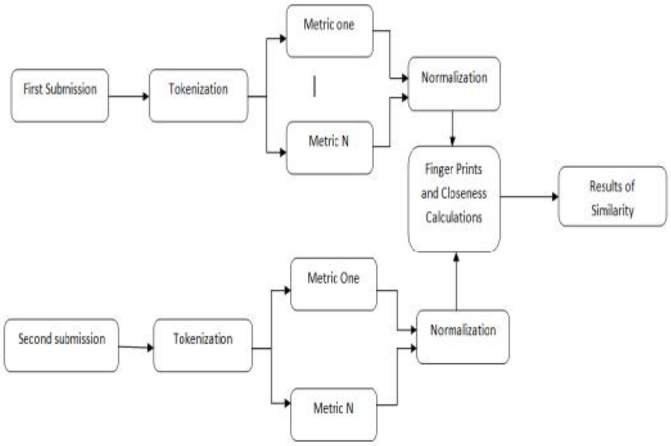


Figure 1: Proposed Model

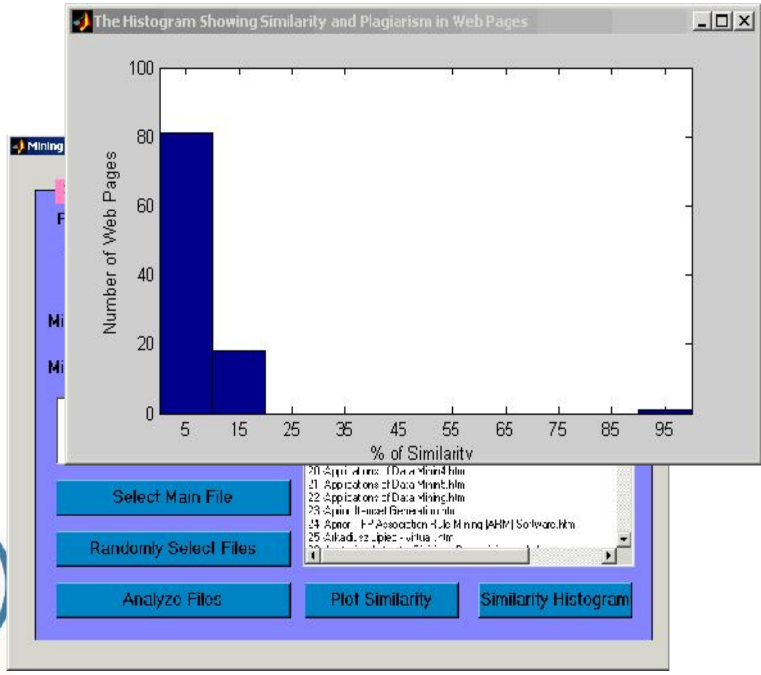
The proposed framework aims to enhance the existing string-matching plagiarism detection approach with similarity analysis techniques. The framework is organised as a five stage approach. The operation of the stages is dependent on the input data, where in some cases not all the stages are required for specific tasks.

- **Stage 1: Pre-processing.** This stage is to prepare the input data, i.e., the entire text collection of suspicious and source texts (corpus), with the language processing techniques which include simple text processing and shallow similarity analysis techniques. This step generalizes the data for feature extraction or comparison in other stages.
- **Stage 2: Similarity comparison.** This stage is to perform pair-wise comparison for all processed texts using a similarity metric. The similarity between text pairs is given by a similarity score, which is then passed on to Stage 3.

- **Stage 3: Filtering.** The similarity scores generated in Stage 2 are used for judging the likelihood for a text-pair to be listed as a candidate pair. The likelihood is usually determined by setting a threshold on the similarity scores. The text-pairs with higher similarity scores are selected for further processing and the rest are discarded. This reduces the search span in the deep linguistic processing stage.
- **Stage 4: Further processing**. Further processing involves the application of deeper language processing techniques, which are computationally expensive to be applied on the whole corpus. When the candidate pairs are retrieved, they are processed by one or more of the modules, generating one or more additional similarity scores.
- **Stage 5: Classification.** The final stage is to give each text pair a classification as Plagiarised or Non-plagiarised. In some cases the Plagiarised class can be further defined in various levels, such as Near Copy, Heavy Revision, or Light Revision. The classification is either done by setting thresholds on the scores from Stage 4, or by using similarity scores generated from various modules in that stage as features in a supervised machine learning classifier.

The proposed framework aims to bring a novel perspective to the traditional pair wise comparison detection approach. The framework is organised as a three-stage approach. As opposed to the traditional external plagiarism detection approach where plagiarised cases and source cases are treated as a pair, the identification of plagiarism direction requires each plagiarised case or source case to be treated on their own. This is done by drawing statistical and linguistic features from each case that can represent rewriting or originality traits. Such features are evaluated individually and in various combinations as supervised machine learning classification or ranking tasks. This sheds light on a number of potential applications, such as first-stage filtering in the traditional plagiarism detection approach, or intrinsic plagiarism detection and authorship identification.

The following screen output shows the input file and some of the files to be searched for web plagiarism. The text boxes in the right side were used to change the web page selection parameters. The tight side text box shows the selected web pages for plagiarism detection.



20 Applications of Data Mining4.htm
21 Applications of Data Minint.htm
22 Applications of Data Mining.htm
23 Apriori-based Generalization.htm
24 Apriori – FP Association Rule Mining (ARM) Software.htm
25 Arkadiusz Lipiec - virtual .htm

Figure 2: The Main Interface in Action

The following output shows the input parameters and the results.

Mining the Web for Plagiarized Web Content

The Test Parameters:
The Specimen File Name : Clustering.htm
The Type of the Files Analyzed : *.htm
The Total Files Used : 100 Nos
The Minimum Size of the Files Used: 1 Kb
The Maximum Size of the Files Used: 1000 Kb

The Test Results:

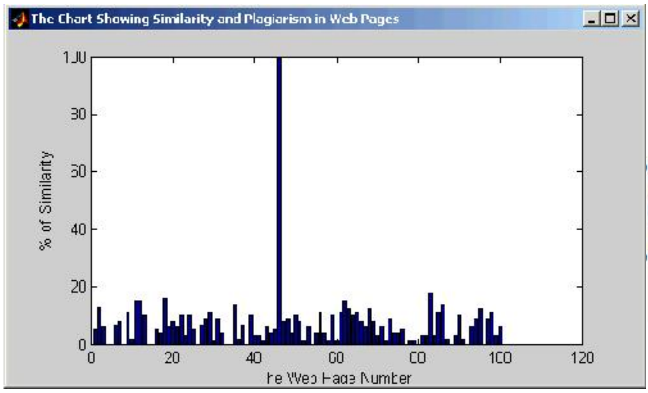
The Percentage of Similarity

81 Files are 5 Percent Similar to the Specimen File
18 Files are 15 Percent Similar to the Specimen File
0 Files are 25 Percent Similar to the Specimen File
0 Files are 35 Percent Similar to the Specimen File
0 Files are 45 Percent Similar to the Specimen File
0 Files are 55 Percent Similar to the Specimen File
0 Files are 65 Percent Similar to the Specimen File
0 Files are 75 Percent Similar to the Specimen File
0 Files are 85 Percent Similar to the Specimen File
1 Files are 95 Percent Similar to the Specimen File

The Total Time Taken for Plagiarism Detection:
3.735000 sec

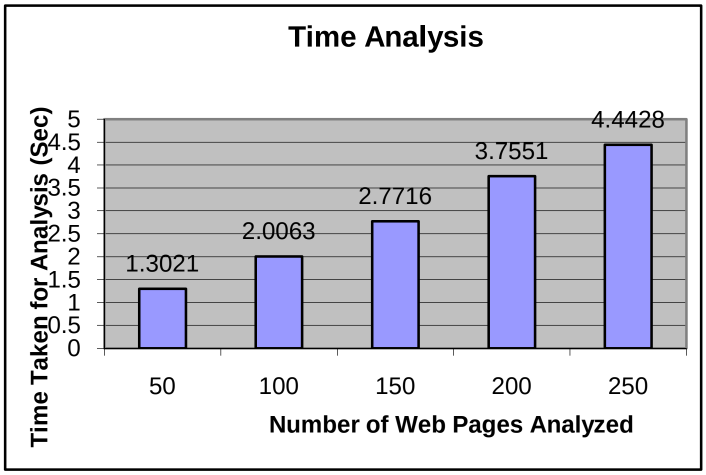
- **Graphical View of Plagiarism Found**

Figure 3: Graphical View of Similarity



The following screen outputs shows the graphical view of similarity of the web documents with respect to the main document.

Figure 4: Timeline Graph for Similarity



The evaluation of our semantic similarity measure using WordNet and Wikipedia resources showed improved performance against baseline statistical methods (stemming, tf/idf weighting and cosine), either supervised or unsupervised approaches are employed for determining the appropriate similarity thresholds. Undeniably, using only the textual information and occurrence statistics is the first step in detecting plagiarism. Mainly because of the complexity of the semantic solution, this preprocessing is necessary in order to narrow the set of suspicious cases.

However, the use of semantic relatedness is necessary to decide for the ambiguous cases of plagiarism. Concerning the scalability of our approach we should mention that it can be embedded in any existing plagiarism detection software in order to improve its results, either it searches for plagiarism in a predefined corpus of essays or it uses the web as a database.

4. CONCLUSION

The standard metric in measuring the amount of shared information between two computer electronic documents was explored to design a Web Page plagiarism detection application in C and Matlab. The proposed Web Page Plagiarism Detection software was successfully implemented and tested with different html files. The results were appreciating. The future versions may have the facility for high-lighting the similar sections for a visual observation. Problems in implementing this kind of interactive observation facilities may be explored in future versions.

The program cannot detect similar repeating sections in the same document. But finding similarity or repetitions in one document is an obvious necessity during evaluating a web page. The possibility of adding this feature may be explored in future.

REFERENCES

1. Broder, A. Z., On the resemblance and containment of documents, Compression and Complexity of Sequences, IEEE Computer Society 1998.
2. Bull, J. et al., Technical Review of Plagiarism Detection Software, Report, Computer-Assisted Assessment Centre, University of Luton 2001.
3. Chandler, A. & Blair, L. Batch Plagiarism Detection with Turnitin, <http://www.comp.lancs.ac.uk/computing/users/angie/plagiarism/batch/Turnitin.htm> Jan 2003
4. CopyCatch Gold, CFL Software Development, <http://www.copycatchgold.com>

5. Duggan F., Advice and Guidance: Data Protection, JISC 2003 <http://online.northumbria.ac.uk/faculties/art/information_studies/Imri/Jiscpas/docs/northumbria/Fiona_workshop.ppt>
6. Gibbon D., Moore R., Winski R., Handbook of Standards and Resources for Spoken Language Systems, Mouton de Gruyter 1997.
7. iParadigms, <http://www.iparadigms.com>
8. JISC plagiarism detection service, <http://www.submit.ac.uk>
9. Lyon C, Malcolm J, and Dickerson B, Detecting short passages of similar text in large document collections, Proceedings of EMNLP (Empirical Methods in Natural Language Processing) 2001.
10. Lyon C, Malcolm J, and Dickerson B, Incremental retrieval of documents relevant to a topic Proceedings of TREC (NIST/DARPA sponsored Text Retrieval Competition) 2002.
11. Manning C. D., and Schutze H., Foundations of Statistical Natural Language Processing, MIT 1999.
12. Ney H., Martin S., Wessel F., Statistical Language Modelling using leaving-one-out. In S Young and G Bloothooft, editors, Corpus Based Methods in Language and Speech Processing. Kluwer Academic Publishers 1997.
13. Spertus E. Mining structural information on the web. In Proceedings of the sixth International World Wide Web Conference,1997. http://decweb.ethz.ch/www6/technical/paper206/paper206.htm.
14. Schechter S,Krishnan M,Smith M.D.Using path profiles to predict HTTP requests.In:Proceedings of the 7th International World Wide Web Conference.Brisbane:Elsevier,1998.http://www7.scu.edu.au/programme/posters/1839/com1839.htm.
15. Data mining:Extending the information warehouse framework.http://www.almaden.ibm.com/cs/quest/paoer/whitepaper.html.
16. ZHOU Hao-feng,ZHU Jian-qiu,ZHU Yang-yong,SHI Bai-le.ARMiner:A Data Mining Tool Based on Association Rules.J.Comput.Sci.&Technol,2002,17(5).594~602
17. http://www.kdnuggets.com/
18. http://www.megaputer.com/