RESEARCH ARTICLE OPEN ACCESS

Plagiarism Detection in Computer Science

S.Munnelli Manohar[1], Mohan Vajjha[2]
Department of Computer Science and Engineering
St. Mary's Group of Institutions
P.N.C & K.R College, Narasaraopet
Guntur
AP-India

ABSTRACT
Recently, the problem of plagiarism is becoming an important issue in many debates in the fields of Education and Technology. The wide use and availability of electronic resources makes it easy for students, authors and even academic people to access and use any piece of information and embed it into his/ her own work without proper citation. The problem is raising in an exponential manner the thing which puts the education process under threat. Several tools are presented to solve the problem of automating plagiarism detection each of which has its own good and bad features, but still the traditional way of plagiarism detection through free text search using search engines is considered an accurate and free way to detect plagiarism with the only disadvantage of being a time consuming method. This research intends to present an alternative to plagiarism detection tools by automating the traditional free search process on search engines to detect plagiarism by intelligently extracting selective parts of text from the file subject to check and pass them to search engine in different forms and processing results in order to come up with a decision of committing plagiarism in a certain degree. The approach used in this paper is to make string comparison of the text with the global *www*, which makes it more comprehensive compared To other plagiarism tools that depend on specific databases.
Keywords:- Plagiarism detection, String matching,

I. INTRODUCTION

Plagiarism is the use of the language and thoughts of another work and the representation of them as one's own original work. Various levels of plagiarism exist in many domains in general and in academic papers in particular. Therefore, diverse efforts are taken to automatically identify plagiarism. In this research, we developed software capable of simple plagiarism detection. We have built a
corpus (C) containing 10,100 academic papers in computer science written in English and two test sets including papers that were randomly chosen from C. A widespread variety of baseline methods has been developed to identify identical or similar papers. Several methods are novel. The experimental results and their analysis show interesting findings. Some of the novel methods are among the best predictive methods.
In light of the explosion in the number of available documents, fast and accurate searching for plagiarism is becoming more needed. Identification of identical and similar documents is becoming very important. Plagiarism is the use of the language and thoughts of another work and the representation of them as one's own original work (Wikipedia, 2010; Library and Information Services, 2010).

Plagiarism can be committed by "recycling" other's work as well as by one's own work (self-plagiarism).
Various levels of plagiarism exist in many domains in general and in academic papers in particular. In addition to the ethical problem, plagiarism in Academics can be illegal if copyright of the previous publication has been transferred to another entity. It is important to mention, that in many cases similar papers are different versions of the same work, e.g., a technical report, a poster paper, a conference paper, a journal paper and a Ph.D. dissertation.

II. PLAGIARISM DETECTION

Plagiarism detection is the process of locating instances of plagiarism within a work or document. The widespread use of computers and the advent of the Internet have made it easier to plagiarize the work of others. Most cases of plagiarism are found in academia, where documents are typically essays or reports. However, plagiarism can be found in virtually any field, including scientific papers, art designs, and source code. Detection of plagiarism can be either manual or software-assisted. Manual detection requires substantial effort and excellent memory, and is impractical in cases where too many documents must be compared, or original documents are not available for comparison. Software-assisted detection allows vast collections of

documents to be compared to each other, making successful detection much more likely. Plagiarism is defined in the 1995 Random House Compact Unabridged Dictionary as the "use or close imitation of the language and thoughts of another author and the representation of them as one's own original work." Self-plagiarism is the reuse of significant, identical, or nearly identical parts of one's own work without citing the original work. In addition to the ethical issue, this phenomenon can be illegal if copyright of the previous work has been transferred to another entity. Usually, self plagiarism is considered to be a serious ethical problem in cases where a publication needs to contain an important portion of a new material, such as in academic papers (Wikipedia, 2010). On the other hand, it is common for researchers to rephrase and republish their research, tailoring it for different academic journals and conference articles, to disseminate their research to the widest possible interested public. However, these researchers must include in each publication a meaningful or an important portion of a new material (Wikipedia, 2010).

III. DETECTION METHODS

The classification of all detection approaches currently in use for computer-assisted plagiarism detection. The approaches are characterized by the type of similarity assessment they undertake: global or local. Global similarity assessment approaches use the characteristics taken from larger parts of the text or the document as a whole to compute similarity, while local methods only examine pre-selected text segments as input.

3.1 Fingerprinting

Fingerprinting is currently the most widely applied approach to plagiarism detection. This method forms representative digests of documents by selecting a set of multiple substrings (n-grams) from them. The sets represent the fingerprints and their elements are called minutiae. A suspicious document is checked for plagiarism by computing its fingerprint and querying minutiae with a precomputed index of fingerprints for all documents of a reference collection. Minutiae matching with those of other documents indicate shared text segments and suggest potential plagiarism if they exceed a chosen similarity threshold. Computational resources and time are limiting factors to fingerprinting, which is why this method typically only compares a subset of minutiae to speed up the computation and allow for checks in very large collection, such as the Internet.

3.2 String Matching

String matching is a prevalent approach used in computer science. When applied to the problem of plagiarism detection, documents are compared for verbatim text overlaps. Numerous methods have been proposed to tackle this task, of which some have been adapted to external plagiarism detection. Checking a suspicious document in this setting requires the computation and storage of efficiently comparable representations for all documents in the reference collection to compare them pair wise. Generally, suffix document models, such as suffix trees or suffix vectors, have been used for this task. Nonetheless, substring matching remains computationally expensive, which makes it a non-viable solution for checking large collections of documents.

3.3 Bag of words

Bag of words analysis represent the adoption of vector space retrieval, a traditional IR concept, to the domain of plagiarism detection. Documents are represented as one or multiple vectors, e.g. for different document parts, which are used for pair wise similarity computations. Similarity computation may then rely on the traditional cosine similarity measure, or on more sophisticated similarity measures.

3.4 Citation analysis

Citation-based plagiarism detection (CbPD) relies on citation analysis, and is the only approach to plagiarism detection that does not rely on the textual similarity. CbPD examines the citation and reference information in texts to identify similar patterns in the citation sequences. As such, this approach is suitable for scientific texts, or other academic documents that contain citations. Citation analysis to detect plagiarism is a relatively young concept. It has not been adopted by commercial software, but a first prototype of a

citation-based plagiarism detection system exists. Similar order and proximity of citations in the examined documents are the main criteria used to compute citation pattern similarities. Citation patterns represent subsequences non-exclusively containing citations shared by the documents compared. Factors, including the absolute number or relative fraction of shared citations in the pattern, as well as the probability that citations co-occur in a document are also considered to quantify the patterns' degree of similarity.

3.5 Stylometry

Stylometry subsumes statistical methods for quantifying an author's unique writing style and is mainly used for authorship attribution or intrinsic CaPD. By constructing and comparing stylometric models for different text segments, passages that are stylistically different from others, hence potentially plagiarized, can be detected.

IV. CONCLUSIONS AND FUTURE WORK

Furthermore, we implemented methods defined for the three thirds of the paper. These methods were combined with CA or CR in various variants. Especially CA and also CR are among the best methods for identification of various levels of plagiarism. In contrast to the best full and selective fingerprint methods, CA and CR check a rather small portion of the papers, and therefore, their run time is much smaller. The success of CA and CR teaches us that most documents that are suspected as simple plagiarized papers include abstracts and references, which have not been significantly changed compared to other documents or vice versa. There is a continuous need for automatic detection of plagiarism due to web influences, and advanced and more complex levels of plagiarism. Therefore, some possible future directions for research are: (1) Developing new kinds of selective fingerprint methods and new combinations of methods to improve detection, (2) Applying this research to larger and/or other corpora, and (3) Dealing with complex kinds of plagiarism, e.g., the use of synonyms, paraphrases, and transpositions of active sentences to passive sentences and vice versa.

REFERENCES

[1] Collberg, C., Kobourov, S., Louie, J., and Slattery, T., 2005. Self-Plagiarism in Computer Science. *Communications of the ACM*, 48(4), pp. 88-94.

[2] Heintze, N., 1996. Scalable Document Fingerprinting. In *Proceedings of the USENIX Workshop on Electronic Commerce*, Oakland California.

[3] Hoad, T. C., and Zobel, J., 2003. Methods for Identifying Versioned and Plagiarised Documents. *Journal of the American Society for Information* Science and Technology, Vol 54(3), pp. 203-215.

[4] IEEE, 2010. Introduction to the Guidelines for Handling Plagiarism Complaints. http://www.ieee.org/publications_standards/publications/rights/plagiarism.html.

[5] Keuskamp, D., and Sliuzas, R., 2007. Plagiarism Prevention or Detection? The Contribution of Text-Matching Software to Education about Academic Integrity. Journal of Academic Language and Learning, Vol 1(1), pp. 91-99.

[6] Lyon, C., Malcolm, J., and Dickerson, B., 2001. Detecting Short Passages of Similar Text in Large Document Collections. In Proceedings of Conference on Empirical Methods in Natural Language Processing, pp. 118-125.

[7] Monostori1, K., Finkel, R., Zaslavsky, A., Hodasz, G., and Patke, M., 2002. Comparison of Overlap Detection Techniques. In Proceedings of the 2002 International Conference on Computational Science, Lecture Notes in Computer Science, vol 2329, pp. 51-60.

[8] Shivakumar, N., and Garcia-Molina, H., 1996. Building a Scalable and Accurate Copy Detection Mechanism. In *Proceedings of the International Conference on Digital Libraries*, pp. 160-168.

[9] Snider, N., and Diab, M., JUNE 2006A. Unsupervised Induction of Modern Standard Arabic Verb Classes. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pp. 153- 156, June 2006.

[10] Sorokina, D., Gehrke, J., Warner, S., Ginsparg, P., 2006. Plagiarism Detection in arXiv. In *Proceedings of Sixth International Conference on Data Mining (ICDM)*, pp. 1070-1075.

[11]Wikipedia, 2010. Plagiarism. http://en.wikipedia.org/wiki/Plagiarism. Witten, I. H., Moffat, A., and Bell, T. C., 1999. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann, second edition.

Munnelli Manohar, M.Tech (C.S) Student of St.Mary's Engineering College. An affiliated college of JNTU Technology University, Kakinada, India. He has done his M.Sc in P.N.C & K.R College, Narasaraopet, Acharya Nagarjuna university, Guntur, AP, India.

Mohan Vajjha, received his MCA from P.N.C & K.R College, Narasaraopet, Acharya Nagarjuna University, Guntur, AP, India.He is currently working as Administrator at Srikari Impetus Solutions. Hyderabad, AP, India