UDC 004.912

PLAGIARISM DETECTION

DZMITRY DZIOKIN, YURY PASTUHOV
Polotsk State University, Belarus

The article is dedicated to plagiarism problem in terms of modern world. Classification of computer-assisted plagiarism detection methods is given.

Introduction. The problem of plagiarism became more serious with the advent of the Internet. Any piece of information that appeared on the Internet just once becomes a public domain. Therefore, it becomes very hard (sometimes even impossible) to observe copyright. It is hard to define original author as well. Impetuous development of the Internet promotes penetration of plagiarism in different spheres of human activity.

Plagiarism is a crime. It often misleads a reader, causes some problems for the author. Mostly, plagiarism is found in universities, where documents are usually a kind of reports or essays. Though, plagiarism can be also found in other fields such as novels, some scientific papers and even source codes.

Plagiarism detection. The process of searching and locating instances of plagiarism within a particular work, paper or document is called plagiarism detection.

Detection of plagiarism can be either manual or software-assisted. Manual detection requires substantial effort and excellent memory, and it 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 thus making successful detection much more likely. [1]

Taking into account the power of present computers, the usage of computer-assisted plagiarism detection becomes very effective. This detection mechanism is an information retrieval task, which is widely supported by specialized information retrieval systems (like Google, Yandex and etc.).

In common words, plagiarism detection mechanism is simple: a suspicious document is compared with a collection of authentic documents. Based on a chosen document model and predefined similarity criteria, the detection task is to retrieve all documents that contain information that is similar to the text in the suspicious document. [2] The figure 1 represents classification of computer-assisted plagiarism detection methods from technical point of view.

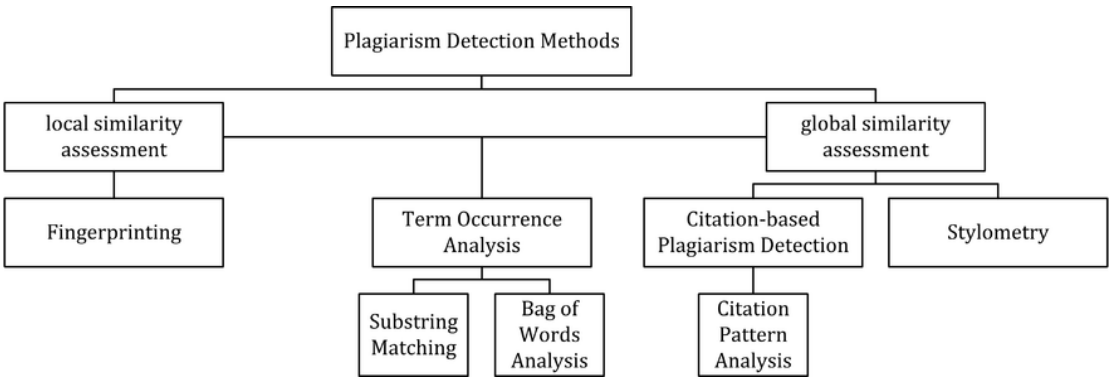

Fig. 1. Classification of computer-assisted plagiarism detection methods

Let's consider these methods:

• Fingerprinting. This is the most widely applied method of plagiarism detection nowadays. 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 [3]. 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 indicates shared text segments and suggests potential plagiarism if a chosen similarity threshold is exceeded. [4]

• String matching. Rather spread method which is used in computer science. Checking a suspicious document in this setting requires the computation and storage of efficiently comparable representations for all doc-

uments in the reference collection. [1] These representations are further compared pairwise. Suffix trees or suffix vectors models are used for this task. However, it is worth noting that this approach is expensive in terms of computing: the algorithm takes 2h comparisons in general, where h is the length of a string where the search is going.

• Bag of words. It is the adoption of vector space retrieval where 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. [5][6][7]

• Citation analysis. This approach is suitable for scientific texts, or other academic documents that contain citations. 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. [8]

• Stylometry uses statistical methods for quantifying an author's unique writing style and is used for authorship attribution.

All detection approaches rely on textual similarity (except for citation pattern analysis). Generally, these mechanisms are analysis stage of the Culwin and Lancaster's four stages of detecting plagiarism. They are collection, analysis, confirmation and investigation. Detection mechanisms depend on some authentic collection of documents, it is very important to have up-to-date and large collection of documents and constantly refresh it.

Conclusion. In the age of information technologies plagiarism has become more actual and turned into a serious problem. The problem of plagiarism is described in this article. Basic approaches of plagiarism detection are reviewed. The most useful and popular methods of computer-assisted plagiarism detection such as fingerprints, string matching, bag of words, citation analysis and stylometry are described.

REFERENCES

1. Plagiarism detection [Electronic resource]. – Mode of access: https://en.wikipedia.org/wiki/Plagiarism_detection. – Date of access: 11.02.2018.
2. Strategies for Retrieving Plagiarized Documents : proceedings 30th Annual International ACM SIGIR Conference, 2007 / Stein, Benno; Meyer zu Eissen, Sven; Potthast, Martin. – P. 825–826.
3. Overview of the 1st International Competition on Plagiarism Detection : PAN09 - 3rd Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse and 1st International Competition on Plagiarism Detection, CEUR Workshop Proceedings, 2009 / Potthast, Martin; Stein, Benno; Eiselt, Andreas; Barrón-Cedeño, Alberto; Rosso, Paolo. – P. 1–9.
4. Copy Detection Mechanisms for Digital Documents, Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, 1995 / Brin, Sergey; Davis, James; Garcia-Molina, Hector. – P. 398–409.
5. CHECK: A Document Plagiarism Detection System", SAC '97: Proceedings of the 1997 ACM symposium on Applied computing, 1997 / Si, Antonio; Leong, Hong Va; Lau, Rynson W. H. – P. 70–77.
6. Dreher, Heinz. Automatic Conceptual Analysis for Plagiarism Detection / Heinz Dreher // Information and Beyond: The Journal of Issues in Informing Science and Information Technology. – 2007. – 4. – P. 601–614.
7. External and Intrinsic Plagiarism Detection Using Vector Space Models : PAN09 - 3rd Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse and 1st International Competition on Plagiarism Detection, CEUR Workshop Proceedings, 2009 / Muhr, Markus; Zechner, Mario; Kern, Roman; Granitzer, Michael. – P. 47–55.
8. Citation Based Plagiarism Detection - A New Approach to Identifying Plagiarized Work Language Independently : Proceedings of the 21st ACM Conference on Hypertext and Hypermedia (HT'10), June 2010 / Gipp, Bela; Beel, Jöran. – P. 273–274.