A Survey on Plagiarism Detection Techniques and Understanding Textual Features

Heena Sherwani[#1], Mitesh Bargadiya[*2]

[#1] *Research Scholar (M.Tech (SS) IV Sem), Department of Computer Science & Engineering*

[*2] *Assistant Professor, Department of Computer Science & Engineering*

[#1, *2] *Vindhya Institute of Technology and Science, Indore (M.P) INDIA)*

Abstract - Web is a powerful tool for human where different types of data and applications are readily available to explore. The unknown or secondary authors copy the information and text from the original document which is readily available on web, leads to copyright violation in web document known as Plagiarism. This paper presents various plagiarism detection techniques. Textual features are needed in plagiarism detection frameworks. The proposed work is intended to survey various plagiarism types, different textual features and plagiarism detection methods to analyse which method detects which plagiarism type. We have also surveyed the plagiarism over watermarked text and plain text document.

Keywords: Plagiarism, textual features, watermarking.

I. INTRODUCTION

In the digital era, various resources are available on the web. Web contains very sensitive and important documents which are readily available to explore. The information distribution through web also involves the risk. A watermark can be a valuable addition to our document. Whether you want to enhance the appearance of the document by adding a seal or image or whether you want to add a text watermark that identifies the document contents as a draft or private information, the watermark feature is a definite asset for intermediate to advanced users. Watermarking discourages counterfeiting. The watermarking is a method to achieve the copyright protection of web contents. Because the multimedia represents several different media such as text, image, video, audio, and graphic objects, and they shows very different characteristics in hiding information inside them so different watermarking algorithms appropriate to each of them should be developed [6].

An effective watermark carries different properties which may change as per application. The properties are [6]:

1. Lustiness: The watermark should be reliable or valid so that it must not degrade easily and should persist under severe conditions.

2. Security: Watermark should maintain the privacy of document from unwanted or unauthenticated sources. It should be able to maintain the secrecy of document.

3. Rapid retrieval: The watermark algorithms must be at full tilt. Easy and fast embedding of documents must be done.

4. Multiple watermarks: The document should be embedded with several watermarks such that each time the image or text downloaded should contain unique watermark.

5. Unambiguity: A watermark should not have redundancy problem regarding document of a rightful author.

II. BACKGROUND

There are no two humans write exactly the same text no matter how similar thoughts and language they are using. Similarly, text written by different authors should be different, except for cited portions. If proper referencing is left or abandoned then problems of plagiarism arise. Alzahrani *et al.* [5] discussed the classification of plagiarism, based on the plagiarist's behaviour into two types:

A. Literal Plagiarism

Literal Plagiarism is the most common practice where plagiarists simply copy and paste the text from web. Except few alterations in the original text document, plagiarists copies entirely word-for-word from source document without direct quotation. The diagram below shows various types of plagiarism [5].

B. Intelligent Plagiarism

Intelligent Plagiarism is another type of plagiarism wherein plagiarists try to cheat readers by changing the contributions of others which appears as their own. In intelligent plagiarism, plagiarists try to hide and change the original work in various intelligent ways which are as follows [5]:

1. Text Manipulation- This plagiarism can confuse the readers by changing most of its appearance and manipulating the text. Paraphrasing and summarizing the text in a shorter form using restructuring, sentence reduction and concept specification are other types of plagiarisms.

2. Translation- Plagiarism can be done by translating the text document which is in one language into another without referencing to original source. Automatic translation and manual translation comes under translated plagiarism.

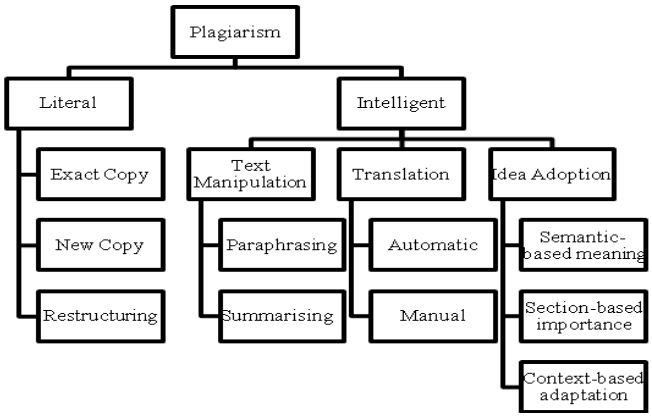

Fig. 2.1 Plagiarism Types

3. Idea Adoption- It is the most critical plagiarism that refers to copy other's idea and use it as their own. It is a major ethical offence adopting to adopt ideas of other. Idea Adoption is a serious academic problem. Idea adoption plagiarism can be committed in following forms:

a) Semantic-based meaning- In this type of idea plagiarism, same idea is expressed in different words using translation, paraphrasing and summarization of text.

b) Section-based importance- Here the most important sections such as findings, conclusions etc are plagiarized. Thus under this type, it plagiarizes the important substantial sections of any document.

c) Context–based meaning- In this type of plagiarism, structure or sequence of ideas is plagiarized from the source document.

PLAGIARISM DETECTION- Plagiarism detection is the process of figuring the document, analysing its data, unveiling the portions that are likely to be plagiarized and bringing similar source documents if available [5]. According to Osman *et al.* Plagiarism detection process has four stages which are collection, analysis, confirmation and investigation [1]. Conference management systems, academic institutions and publishers have started using plagiarism detectors such as CrossCheck, Turnitin, WCopyFind etc. Among these detectors Turnitin is very popular [4].

PLAGIARISM DETECTION TASKS- plagiarism detection is divided into two main tasks [1], [5] which are:

1. Extrinsic Plagiarism Detection- It detects plagiarism on a reference to one or more source documents. It uses computer's capability to search the plagiarized document in a similar source documents. Many researches have been undertaken for extrinsic plagiarism detection. In this task the various source documents and a query document is taken which passes through retrieval model later a feature based analysis is done and comparison is done between different units. Now processing unit combines theses units into passages to present the results to human which may decide that plagiarism is committed or not.

2. Intrinsic Plagiarism Detection- Intrinsic plagiarism detection is carried out by looking into query document in isolation. It uses human's capability to detect plagiarism via writing style variations [3]. Stamatatos *et al.* has presented new method to quantify style variations using character n-gram profiles and a style change function [3]. Thus intrinsic plagiarism detection deals where no reference is given and style inconsistencies are presented in a document.

TEXTUAL FEATURES- Textual features are used to characterize documents before plagiarism detection methods are applied. Textual features quantify documents and are needed in various plagiarism detection frameworks which are as follows:

A. Textual Features for Extrinsic Plagiarism Detection- To quantify documents in extrinsic plagiarism detection, various textual features are discussed below with detailed description [5]:

1) Lexical Features: In any document, character and word are the simplest form to represent. Lexical features are operated at character or word level. Character *n*-gram representation is used in a document d to represent a sequence of characters. Similarly, word *n*-gram representation is a collection of words in any document. *CNG* and *WNG* are known as *shingles* or *fingerprints.* The process of creating fingerprints is called *fingerprinting* or *shingling.* Various tools and resources such as Tokenizer, Stemmer, Lemmatizer are used to break document into its basic form [8].

2) Syntactic Features: They are characterized by sentence based representation. Here text is splitted using end delimiters such as full stop, question mark etc. After splitting the text into sentences, part of speech (POS) structure can be created using POS tagger. Similarly text chunk is used to characterize bigger text and chunker is generated by windowing. Tools and resources used are POS tagger, text chunker, Sentence splitter, Partial parser etc.

3) Semantic Features: Semantic features helps in plagiarism detection by providing insights into the meaning of text so that we can compare text semantically using semantic dependencies and POS tagging. Thesaurus, Word-Net etc helps to find the synonyms, hyponyms and antonyms.

4) Structural Features: Structural features for plagiarism detection deals with tree organizations of documents. Document is the collection of paragraphs and similar paragraphs constitute a block having similar semantics. These blocks later form sections and sub-sections. This semi structured document such as journal papers etc quantified by structural features. Structural features can be divided into two types which are *block-specific* and *content-specific structured features* [5]. In block specific structured features, web pages based documents were considered as blocks. Later HTML tags are used to segment web pages. The paragraphs are grouped into pages where a new paragraph is added to each page until maximum count is reached otherwise new page is created. Content-specific structured features encode the document features into semantically related blocks. Content-specific features with flat features can get better gist of concepts thus provides better plagiarism detection.

B. Textual Features for Intrinsic Plagiarism Detection- Every author has its own writing skills and style. Intrinsic plagiarism detection deals with stylometric features which quantify the writing style and its variations. Intrinsic plagiarism detection also uses lexical features operate at character level, syntactic features operate at sentence level and uses POS, semantic features quantify semantic dependencies and application based features which includes language specific and content-specific keywords [5].

PLAGIARISM DETECTION METHODS- Various research works have undertaken and as a result gives different plagiarism detection techniques which are as below [1], [5]:

1) Character-Based Methods: According to Alzahrani *et al.* [5] character based techniques helps in plagiarism detection using character *n*-gram and word *n*-gram features. The suspicious document is compared against the candidate document on string basis. String matching can be exactly similar or approximate. *Exact* string matching means two strings say x and y have exactly the same letters. While in *approximate* string matching some degree of letter resembles and thus different operations like *Insertion, Deletion, Substitution* and *Transposition* are applied. This string matching is also called as *fingerprint.*

2) Vector-Based Methods: Like character-based method, vector-based method relies on tokens instead of strings. Tokens are compared and similarity is calculated using vector similarity coefficient, Dice's, Euclidean, Cosine, or Manhattan coefficients. Due to simplicity and ease, Cosine with other metrics was efficient for plagiarism detection.

3) Syntax-Based Methods: Syntactical features are used to find the similarity between texts. This syntactical features contains POS that is part of speech tagger which gauge the similarity between texts. The similarity using POS tags shows that two texts are having similar syntactical features and using LCS algorithm [1] it detects plagiarism.

4) Semantic-Based Methods: A sentence is a group of words arranged in different order. Two sentences can be semantically same but having different order of words. Thus resemblance between two sentences can be derived through *word* similarity and *order* similarity [5]. Osman *et al.*, [2] have introduced effective plagiarism detection technique based on Semantic Role Labeling. SRL analyzed each sentences in a text and generated a value for each argument. Finally, the generated values for each argument show their behavior and also show that every argument behavior does not affect the plagiarism detection process [2].

5) Fuzzy-Based Methods: The concept of "fuzzy" in plagiarism detection is related with vagueness of similarity of words in a sentence of any text. The similarity may vary from zero to one which means degree of similarity between words in a document match exactly in case of one and match nothing in case of zero. While there may be some degree of similarity is shown by fuzzy set. A *term-to-term correlation matrix* is constructed which shows the similarity between different words and helps to form fuzzy set. Fuzzy-based methods found to be very effective up to some extent.

6) Structural-Based Methods: All the above methods use flat features representations such as lexical features, syntactical features and semantic features of a text in a document. Hierarchical feature or tree feature representation uses contextual information and shows contextual similarity throughout the document or section or sub-section. Tree structured representation is very effective in plagiarism detection of web pages and web related documents.

7) Stylometric-Based Methods: Every author has their own writing skills and vocabulary. Author employs patterns to construct sentences having different writing style. Stylometric-based methods are used in intrinsic plagiarism detection frameworks where plagiarism is detected by quantified writing style variation.

8) Cross-Lingual Plagiarism Detection Methods: Cross lingual methods using cross language features measures the similarity between sections of query document and sections of the candidate document. It uses various cross-lingual syntax-based methods, cross-lingual semantic-based methods, cross-lingual dictionary-based methods and statistics-based methods [5].

III. REVIEW

In the following Table 1 contains the literature review, along with their contributions in the Plagiarism Detection techniques and antiplagiarism tools.

TABLE I : SUMMARY OF VARIOUS AUTHOR'S CONTRIBUTION IN PLAGIARISM DETECTION

S no.	Year	Published By	Paper Title	Remark
1	2012	A.H. Osman *et al.*	"Survey of Text Plagiarism Detection"	1. This paper has significantly reflected a novel text plagiarism detection technique called citation-based methods. 2. Author has described the four plagiarism detection process stages as collection, analysis, confirmation and investigation.
2	2011	A. H. Osman e*t al.*	"An Improved Plagiarism Detection Scheme Based on Semantic Role Labeling"	1. In this paper, author has developed an improved plagiarism detection technique on semantic role labeling in which SRL employed and compared text based on semantic allocation for every term inside a sentence.
3	2008	Efstathios Stamatatos *et al.*	"Intrinsic Plagiarism Detection Using Character *n*-gram Profiles"	1. In this paper, author has presented a new method to quantify style variation within a document using style change function and character *n*-gram profile. 2. Author proposed a set of heuristic ruled to detect plagiarism.
4	2006	Hermann Maurer *et al.*	"Plagiarism- A Survey"	1. This paper shows a detailed survey on various plagiarism detection tools and services. 2. Various web based tools such as Turnitin, SafeAssignment, docoloc, Moss etc are investigated and compared effectively.
5	2012	Alzahrani *et al*	"Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods"	1. This paper showed state-of-the-art techniques for plagiarism detection, different frameworks and textual features with resources and tools. 2. Author suggested two methods semantic-based and fuzzy-based for idea plagiarism detection. 3. Author also proposed the use of structural-based methods are efficient for section and context-based idea plagiarism detection.
	2014	Sunanda Datta *et al.*	"Data Authentication using Digital Watermarking"	1. In this paper, author has explained the digital watermarking concepts along with different watermarking methods and which method is more suitable to be used. 2. Various types, processes and contributions of other authors in the field of watermarking have been presented clearly.
7	2012	Vanessa Wei Feng *et al.*	"Text-level Discourse Parsing with Rich Linguistic Features"	1. In this paper, author has developed an RST-style text-level discourse parser and incorporated their own rich linguistic features.
8	2014	Zdenek Ceskal *et al.*	"Multilingual Plagiarism Detection"	1. This paper describes a novel approach and a new method called MLPlag for multilingual plagiarism detection. This method has proved to be accurate and has given promising results. 2. Experiments are performed on both monolingual and multilingual corpora and results are shown in this paper.

IV. CONCLUSION

This paper includes the survey of plagiarism detection techniques and various textual features for capturing different types of plagiarism. Character-based methods, vector-based methods and syntax-based methods are suitable for detecting literal plagiarism. Semantic-based and fuzzy-based methods are more reliable as they incorporate semantic features. Both these methods received less attention because of time complexity of algorithms which makes them impractical for real use. Stylometric methods are used in intrinsic plagiarism detection framework in which source document is unavailable and leaves no evidence of plagiarism which is important aspect for human. Structure-based methods are used for tree structure representation of document and thus combined with vector-based methods detects plagiarism more efficiently. However all these methods cater literal plagiarism but idea plagiarism needs more research and more effective approach.

REFERENCES

[1] Ahmed Hamza Osman, Naomie Salim, and Albaraa Abuobieda, ”Survey of Text Plagiarism Detection,” Computer Engineering and Applications Vol. 1, No. 1, June 2012.

[2] A. H. Osman, *et al.*, "An Improved Plagiarism Detection Scheme Based on Semantic Role Labeling," Applied Soft Computing, 2011.

[3] Efstathios Stamatatos, “Intrinsic Plagiarism Detection Using Character *n*-gram Profiles,” University of the Aegean, 2008.

[4] Hermann Maurer, Frank Kappe and Bilal Zaka, “Plagiarism-A Survey,” Journal of Universal Computer Science, vol. 12, no. 8 (2006).

[5] Salha M. Alzahrani, Naomie Salim, and Ajith Abraham, “Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods,” IEEE Transactions On Systems, Man and Cybernetics Part-C, Vol.42, No.2, March 2012.

[6] Sunanda Datta, Dr. Asoke Nath, “Data Authentication Using Digital Watermarking,” Vol.2,Issue 14,December 2014, ISSN: 2321-7782.

[7] Vanessa Wei Feng, Graeme Hirst, “Text-level Discourse Parsing with Rich Linguistic Features,” Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 60–68, Jeju, Republic of Korea, 8-14 July 2012.

[8] Zdenek Ceska1, Michal Toman1 and Karel Jezek, “Multilingual Plagiarism Detection,” University of West Bohemia, oct. 2014.