IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 1, Issue 1, March, 2013
ISSN: 2320 - 8791
www.ijreat.org

Intelligent Plagiarism Detection

Ramya L [1] & Mrs Venkatalakshmi R [2]

[1]Department of Information Technology, SRM University, Chennai, Tamil Nadu, India.

[2]Assistant professor, Department of Information Technology, SRM University, Chennai, Tamil Nadu, India.

Abstract
Plagiarism can be of many different natures, ranging from copying texts to adopting ideas, without giving credit to its originator. This paper presents a new taxonomy of plagiarism that highlights differences between literal plagiarism and intelligent plagiarism, from the plagiarist's behavioral point of view. The taxonomy supports deep understanding of different linguistic patterns in committing plagiarism, for example, changing texts into semantically equivalent but with different words and organization, shortening texts with concept generalization and specification, and adopting ideas and important contributions of others. Different textual features that characterize different plagiarism types are discussed. Systematic frameworks and methods of monolingual, extrinsic, intrinsic, and cross-lingual plagiarism detection are surveyed and correlated with plagiarism types, which are listed in the taxonomy. We conduct extensive study of state-of-the-art techniques for plagiarism detection, including character *n*-gram-based (CNG), syntax-based(SYN), semantic-based (SEM), structuralbased (STRUC), stylometric-based (STYLE), and cross-lingual techniques (CROSS).Our study corroborates that existing systems for plagiarism detection focus on copying text but fail to detect intelligent plagiarism when ideas are presented in different words.

Index Terms—Linguistic patterns, plagiarism, plagiarism detection, taxonomy, textual features.

1. INTRODUCTION

The problem of plagiarism has recently increased because of the digital era of resources available on theWorldWide Web. Plagiarism detection in natural languages by statistical or computerized methods has started since the 1990s, which is pioneered by the studies of copy detection mechanisms in digital documents . Earlier than plagiarism detection in natural languages, code clones and software misuse detection has started since the 1970s by the studies to detect programming code plagiarism in Pascal and C . Algorithms of plagiarism detection in natural languages and programming ferent textual features and diverse methods of detection, while the latter mainly focuses on keeping track of metrics, such as number of lines, variables, statements, subprograms, calls to subprograms, and other parameters. During the last decade, research on automated plagiarism detection in natural languages has actively evolved, which takes the advantage of recent developments in related fields like information retrieval (IR), crosslanguage information retrieval (CLIR), natural language processing, computational linguistics, artificial intelligence, and soft computing. In this paper, a survey of recent advances in the area of automated plagiarism detection in text documents is presented, which started roughly in 2005, unless it is noteworthy to state a research prior than that.

2. PROPOSED SYSTEM

This paper brings patterns of plagiarism together with textual features for characterization of each pattern and computerized methods for detection. The contributions of this paper can be summarized as follows: First, different kinds of plagiarism are organized into a taxonomy that is derived from a qualitative study and recent literatures about the plagiarism concept. The taxonomy is supported by various plagiarism patterns (i.e., examples) from available corpora for plagiarism. Second, different textual features are illustrated to represent text documents for the purpose of plagiarism detection. Third, methods of candidate retrieval and plagiarism detection are surveyed, and correlated with plagiarism types, which are listed in the taxonomy.

3.PLAGIARISM TAXONOMY AND PATTERNS:

There are no two humans, no matter what languages they use and how similar thoughts they have, write exactly the same text. Thus, written text, which is stemmed from different authorsshould be different, to some extent, except for cited portions. If proper referencing is abandoned, problems of plagiarism andintellectual property arise.The existence of academic dishonestyproblems has led most, if not all, academic institutions andpublishers to set regulations against the offence. Borrowed contentof any form require directly or indirectly quoting, in-textreferencing, and citing the original author in the list of references

3.1. Literal Plagiarism

Literal plagiarism is a common and major practice wherein plagiarists do not spend much time in hiding the academic crime they committed. For example, they simply copy and paste the text from the Internet. Aside from few alterations in the original text (marked as underlined), Fig. 3 shows a pattern of text taken entirelyword-for-word from the source without direct quotation

meaning requires citations around the borrowed ideas and citing the original author.

Besides paraphrasing, summarizing the text in a shorter form using sentence reduction, combination, restructuring, paraphrasing, concept generalization, and concept specification is another form of plagiarism unless it is cited properly. Fig. 5 shows that some sentences are combined and restructured, some phrases are syntactically changed, sentences are reduced by liminating underlined text in the original text, and synonyms of some words are used in the summary. Although much of the text is changed and fewer phrases are left in the summary, citation and attribution are still required.

3.2. Intelligent Plagiarism

Intelligent plagiarism is a serious academic Borrowing a few words, but no original ideas, to improve thequality of the English, especially by nonnatives, should not be considered plagiarism . The qualitative study showed that university professors can suspect or detect different types of idea plagiarism using their own expertise. However, computerized solutions for the purpose of detecting idea plagiarism are highly needed, since it is crucial to judge the *quality* of different academic work, including theses, dissertations, journal papers, conference proceedings, essays, and assignments. Idea plagiarism can be classified into three types yet with fuzzy boundaries: semantic-based meaning, section-based importance, and context-based plagiarismdishonesty wherein plagiarists try to deceive readers by changing the contributions of others to appear as their own. Intelligent plagiarists try to hide, obfuscate, and change the original work in various intelligent ways, including text manipulation, translation, and idea adoption.

3.2.1) Text Manipulation: Plagiarism can be obfuscated by manipulating the text and changing most of its appearance. Fig. 2 exemplifies lexical and syntactical paraphrasing, where underlined words are replaced with synonyms/antonyms, and short phrases are inserted to change the appearance, but not the idea, of the text. Paraphrasing while retaining the semantic meaning requires citations around the borrowed ideas and citing the original author.

3.2.2) Translation: Obfuscation can also be done by translating the text from one language to another without proper referencing to the original source. Translated plagiarism includes automatic translation (e.g.,Google translator) and manual translation (e.g., by people who speak both languages). Back translated plagiarism is another (easier) form of paraphrasing by automatically translating a text from one language to another and retranslate it back to the first one fig 3 shows an example of text translated from English to French and back from French to English. It is obvious that the retranslated text may have poor English, but plagiarists could use spell checkers and other text manipulations to obfuscate plagiarism.

3.2.3) Idea Adoption: Idea adoption is the most serious plagiarism that refers to the use of other's ideas, such as results, contributions, findings, and conclusions, without citing the original source of ideas. It is a major offence to steal ideas of others, which is a real academic problem that needs to be investigated.

"Copying a few sentences that contain no original idea (e.g., in the introduction) is of marginal importance compared to stealing the ideas of others.

Figure:1

Original: The definition of term includes single words, keywords, or longer phrases [active voice]. If the words are chosen to be the terms, the dimensionality of the vector is the number of words in the vocabulary.
Plagiarized: Keywords, single words or longer phrases are included in the definition of the term [active-passive conversion]. If the words are chosen to be the terms, the dimensionality of the vector is the number of words in the vocabulary.

Figure 2:

Original: If a term occurs in the document, its value in the vector is non-zero. Several different ways of computing these values, also known as (term) weights, have been developed. One of the best known schemes is tf-idf weighting.
Plagiarized: If a token[synonym] appears[synonym] in the text[synonym] then its value in the vector is non-zero. A couple of[synonym] different algorithms [synonym] of calculating[synonym] these values, also known as (term) weights, have been created. One of the excellent[synonym] known methods[synonym] is called[synonym] tf-idf weighting.

Figure 3:

Original (English): I have a dream that one day this nation will rise up and live out the true meaning of its creed: "We hold these truths to be self-evident, that all men are created equal."
Translated (French): J'ai un rêve que pendant un jour cette nation montera vers le haut et vivra dehors la signification vraie de sa foi : " ; Nous tenons ces vérités pour pour évidents en soi, ce tous les hommes sommes equal." créés
Retranslated (English): I have a dream that during one day this nation will go up to the top and live outside the true significance of its faith: " ; We hold these truths for obvious in oneself, this all men naps equal."

4. PLAGIARISM DETECTION TASKS

Plagiarism detection is divided into two formal tasks: extrinsic and intrinsic . Extrinsic plagiarism detection evaluates plagiarism in accordance to one or more source documents. Intrinsic plagiarism detection, on the other hand, evaluates instances of plagiarism by looking into the suspicious/query document in isolation.The first one utilizes the computer's capability in searching large text collection and retrieving possible sources for plagiarism, whereas the second one simulates the human's ability to catch plagiarism via writing style variations.

4.1. Extrinsic Plagiarism Detection

Extrinsic plagiarism detection is a method of comparing a suspicious document against a set of source collection whereby several text features are used to suspect plagiarism. .

4.2 *Intrinsic Plagiarism Detection*

Intrinsic plagiarism detection, authorship verification, and authorship attribution are three similar tasks yet with different end goals. In all of them, writing style is quantified and/or feature complexity is analyzed. The different end goals of these tasks are

1) to suspect plagiarism in the intrinsic plagiarism detection;
2) to verify whether the text stems from a specific author or not in the authorship verification; and
3) to attribute the text to authors in the authorship attribution.

"Intrinsic plagiarism aims at identifying potential plagiarism by analyzing a document with respect to undeclared changes in writing style. Authorship verification aims at determining whether or not a text with doubtful authorship is from an author A, given some writing examples of A, while authorship attribution aims at attributing a document d of unknown authorship, given a set D of candidate authors with writing examples" That is, intrinsic plagiarism detection can be viewed as the generalization of authorship verification and attribution because intrinsic plagiarism detection analyses the query document in isolation, while authorship analysis problems analyze a document with respect to a set of writing examples of a specific author in authorship verification or a set of candidate authors writing examples in authorship attribution. Many researchworks have been conducted to tackle the task of intrinsic plagiarism detection .

4.3. Plagiarism Detection Languages

Plagiarism detection can be classified into monolingual and cross-lingual based on language homogeneity or heterogeneity of the textual documents being compared

1) Monolingual Plagiarism Detection: Monolingual plagiarism detection deals with the automatic identification and extraction of plagiarism in a homogeneous language setting, e.g., English–English plagiarism. Most of the plagiarism detection systems have been developed for monolingual detection, which is divided into two former tasks, extrinsic and intrinsic, as discussedearlier.

2) Cross-Lingual Plagiarism Detection: Cross-language (or multilingual) plagiarism detection deals with the automatic identification and extraction of plagiarism in a multilingual setting, e.g., English–Arabic plagiarism. Research on crosslingualplagiarism detection has attracted attention in recentfew years thus focusing on text similarity computation across languages.

TABLE 1:TYPES OF STYLOMETRIC FEATURES WITH COMPUTATIONAL TOOLS REQUIRED FOR THEIR MEASUREMENT

-	Examples	Tools and Resources	Ref.
Lexical features (Character-based)	Frequency of characters	-	[39]
	Character types (letters, digits, punctuations, etc.)	Character dictionaries	
	Frequency of special characters (e.g. !,& ,etc.)		
	Character n-grams (fixed-length) frequency	Chunker	[40, 41]
	Character n-grams (variable-length) frequency	Feature selector	
	Compression methods	Text compression tool	[22]
Lexical features (Word-based)	Token-based : - Average word length - Average sentence length - Average syllables per word	Tokenizer, [Sentence splitter]	[39, 49]
	Vocabulary richness - Type-token ratio (i.e. total unique vocabulary/total tokens) - Hapax legomena/dislegomena	Tokenizer	[39, 49-51]
	Frequency of words	Tokenizer, [Stemmer, Lemmatizer]	[40, 49]
	Frequency of function words	Tokenizer, Special dictionaries	[39, 40, 49, 56, 57]
	Word n-grams frequency	Tokenizer	[59]
	Averaged word frequency class	Tokenizer, [Stemmer, Lemmatizer]	[33]
	Lexical Errors - Spelling errors (e.g. letter omissions and insertions) - Formatting errors (e.g. all caps letters)	Tokenizer, Orthographic spell checker	[21, 57]
Syntactic features	Part-of-speech	Tokenizer, Sentence splitter, POS tagger	[22, 40]
	Part-of-speech n-gram frequency		[40, 57]
	Chunks	Tokenizer, Sentence splitter, [POS tagger],	[41, 66]
	Sentence and phrase structure	Tokenizer, Sentence splitter, POS tagger, Partial parser	[67]
	Rewrite rules frequencies	Tokenizer, Sentence splitter, POS tagger, Full parser	[68, 69]
	Syntactic Errors - Sentence fragments - Run-on sentences - Mismatched tense	Tokenizer, Sentence splitter, Syntactic spell checker	[57]
Semantic features	Synonyms, hypernyms, etc.	Tokenizer, [POS tagger], Thesaurus	[70]
	Semantic dependencies	Tokenizer, Sentence splitter, POS tagger, Partial parser, Semantic parser	[69]
	Functional	Tokenizer, Sentence splitter, POS tagger, Thesaurus, Specialised dictionaries	[72]
Application-specific	Structural - Average paragraph length - Indentation - Use of greetings and farewells - Use of signatures	HTML parser, Specialised parsers	[22, 39]
	Content-specific keywords	Tokenizer, [Stemmer, Lemmatizer], Specialised dictionaries	[39]
	Language-specific		[75, 76]

Optional tools are included in brackets.

5. PLAGIARISM TYPES, FEATURES AND METHODS: WHICHMETHOD DETECTS WHICH PLAGIARISM?

The taxonomy of plagiarism (see Fig. 2) illustrates different types of plagiarism on the basis of the way the offender (purposely) changes the plagiarized text. Plagiarism is categorized into literal plagiarism (refers to copying the text nearly as it is) and intelligent plagiarism (refers to illegal practices of changing texts to hide the offence including restructuring, paraphrasing, summarizing, and translating). Adoption of (embracing as your
own) ideas of other is a type of intelligent plagiarism, where a plagiarist deliberately 1) chooses texts that convey creative ideas, contributions, results, findings, and methods of solving problems; 2) obfuscates how these ideas were written; and
3)embeds them within another work without giving credit to the source of ideas. We categorize idea plagiarism, based on its occurrence within the document, into three levels: the lowest is the *semantic-based meaning* at the paragraph (or sentence) level, the intermediate is the *section-based importance* at the section level, and the top (or holistic) is the *context-based adaptation*, which is based on ideas structure in the document.
Textual features are essential to capture different types of plagiarism. Implementing rich feature structures should lead to the detection of more types of plagiarism, if a proper method and similarity measure are used as well. *Flat-feature* extraction includes *lexical*, *syntactic*, and *semantic* features,but does not account contextual information of the document. *Structural-feature* (or *tree-feature*) extraction, on the other hand, takes into account the way words are distributed throughout the document. We categorize structural features into blockspecific which encodes the document as hierarchical blocks (*document-page-paragraph* or *document-paragraphsentence*), and content-specific, which encodes the content as semantic-related structure (*document-section-paragraph* or *class-concept-chunk*). The latter, combined with *flat* features, is suitable to capture a document's semantics and get the gist of its concepts. Besides, we can drill down or roll up through the *tree* representation to detect more plagiarism patterns. Many plagiarism detection methods focus on copying text with/without minor modification of the words and grammar. In fact, most of the existing systems fail to detect plagiarism by paraphrasing the text, by summarizing the text but retaining the same idea, or by stealing ideas and contributions of others. This is why most of the current methods do not account the overlap when a plagiarized text is presented in different words

.

6. CONCLUSION

Current antiplagiarism tools for educational institutions, academicians, and publishers "can pinpoint only word-for-word plagiarism and only some instances of it" and do not cateradopting ideas of others . In fact, idea plagiarism is awfully more successful in the academic world than other types because academicians may not have sufficient time to track their own ideas, and publishers may not be well-equipped to check where the contributions and results come from . As plagiarists become increasingly more sophisticated, idea plagiarism is a key academic problem and should be addressed in future research.

detect *semantic-based meaning* idea plagiarism at the paragraph. We also propose the use of structural features ancontextual information with efficient STRUC-based methods to detect *section-based importance* and *context-based adaptation* idea plagiarism.

6. REFERENCES:

[1] Z. Ceska, "The future of copy detection techniques," in *Proc. YRCAS*, Pilsen, Czech Republic, pp. 5–10.
[2] D. I. Holmes, "The evolution of stylometry in humanities scholarship," *Lit Linguist Comput.*, vol. 13, pp. 111–117, 1998.
[3] O. deVel, A. Anderson, M. Corney, and G. Mohay, "Mining e-mail content for author identification forensics," *SIGMOD Rec.*, vol. 30, pp. 55–64, 2001.
[4] F. J. Tweedie and R. H. Baayen, "How variable may a constant be? Measures
of lexical richness in perspective," *Comput. Humanities*, vol. 32, pp. 323–352, 1998.
[5] P. Clough, "Plagiarism in natural and programming languages: An overview of current tools and technologies," Dept. Comput. Sci., Univ. Sheffield, U.K., Tech. Rep. CS-00-05, 2000.
[6] P. Clough, (2003) Old and new challenges in automatic plagiarism detection.
National UK Plagiarism Advisory Service. [Online]. Available: http://ir.shef.ac.uk/cloughie/papers/pas_plagiarism
[7] H. Maurer, F. Kappe, and B. Zaka, "Plagiarism—A survey," *J. Univ.*
Comput. Sci., vol. 12, pp. 1050–1084, 2006.
[8] L. Romans, G. Vita, and G. Janis, "Computer-based plagiarism detection
methods and tools: An overview," presented at the Int. Conf. Comput. Syst. Technol., Rousse, Bulgaria, 2007.
[9] S. Argamon, A. Marin, and S. S. Stein, "Style mining of electronic messages for multiple authorship discrimination: First results," presented
at the 9th ACM SIGKDD Int. Conf. Know. Discovery Data Mining, Washington, DC, 2003.
[10] M. Koppel and J. Schler, "Exploiting stylistic idiosyncrasies for authorship attribution," in *Proc. IJCAI Workshop on Computat. Approaches*
Style Anal. Synth., Acapulco, Mexico, 2003,
[10] S. Alzahrani, "Plagiarism auto-detection in arabic scripts using statement-based fingerprints matching and fuzzy-set information