Plagiarism: Taxonomy, Tools and Detection Techniques

Dhruba K Bhattacharyya, FIETE

Tezpur University

dkb@tezu.ernet.in

October 28, 2016

Overview

Introduction I

- Plagiarism is the presentation of another's words, work or idea as one's own[4]. It has two components[4]
  1. Taking the words, work or ideas from some source(s).
  2. Presenting it without acknowledgments of the source(s) from where words, works or ideas are taken.
- There are mainly two types of plagiarisms typically found to occur[8]
  1. Textual plagiarism
  2. Source code plagiarism.

[4] Melissa S Anderson and Nicholas H Steneck. The problem of plagiarism. In Urologic Oncology: Seminars and Original Investigations, volume 29, pages 90-94. Elsevier, 2011.

[8] Netra Charya, Kushagra Doshi, Smit Bawkar, and Radha Shankarmani. Intrinsic plagiarism detection in digital data 2(3), 23-30, 2015.

Introduction II

Plagiarism can occur between two same or two different natural languages. Based on language homogeneity or heterogeneity of the textual documents being compared, the plagiarism detection can be divided into two basic types[3].

1. *Monolingual Plagiarism Detection:* This type of detection deals with homogeneous language settings e.g., English-English
2. *Cross-Lingual Plagiarism Detection:* This detection approach is able to perform in heterogeneous language settings e.g., English-Chinese.

[3] 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 (Applications and Reviews), 42(2):133-149, 2012.

Introduction III

There are mainly two types of plagiarism detection approaches available based on whether external resources or references are used or not during plagiarism detection[3].

a *Intrinsic plagiarism detection :* Where no external references are used.

b *Extrinsic plagiarism detection :* Where external references are used.

[3] 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 (Applications and Reviews), 42(2):133-149, 2012.

Contributions I

- It reports a comprehensive and systematic survey on a large number of methods of plagiarism detection and analyzes their pros and cons.
- It includes discussion on a large number of tools on plagiarism detection and reports their features. It also compares these tools based on a set of crucial parameters.
- It introduces the framework of a hybrid detection approach to identify text similarity based on both intrinsic and extrinsic plagiarism.
- Finally, in includes a list of issues and research challenges.

Definition of plagiarism

As stated in [4], plagiarism can be defined as an appropriation of the ideas, words, process or results of another person without proper acknowledgment, credit or citation. Plagiarism can appear in a research article or program in following ways:

a Claiming another person's work as your own.

b Use of another person's work without giving credit.

c Majority of someone's contribution as your own, whether credit is given or not.

d Restructuring the other works and claiming as your own work.

e providing wrong acknowledgment of other works in your work.

[4] Melissa S Anderson and Nicholas H Steneck. The problem of plagiarism. In Urologic Oncology: Seminars and Original Investigations, volume 29, pages 90-94. Elsevier, 2011.

Types of plagiarism

Plagiarism can appear in different forms in a document, work, production or program. Two basic types of plagiarisms[2].

1. *Textual plagiarism:* It is commonly seen in education and research. Figure 1 (a) shows an example of textual plagiarism.
2. *Source Code plagiarism:* Here, codes written by others are copied or reused or modified or converted a part of codes and claimed as one's own. Figure 1 (b) shows the example an example of source code.

[2] Asim M El Tahir Ali, Hussam M Dahwa Abdulla, and Vaclav Snasel. Overview and comparison of plagiarism detection tools. In DATESO, pages 161-172. Citeseer, 2011..

Example of plagiarism

The nearest-neighbour based outlier mining technique is able to detect a plagiarized text segment.(*Active voice*)
A plagiarized text segment is detected by the nearest-neighbour based outlier mining technique.(*passive voice*)

(a)

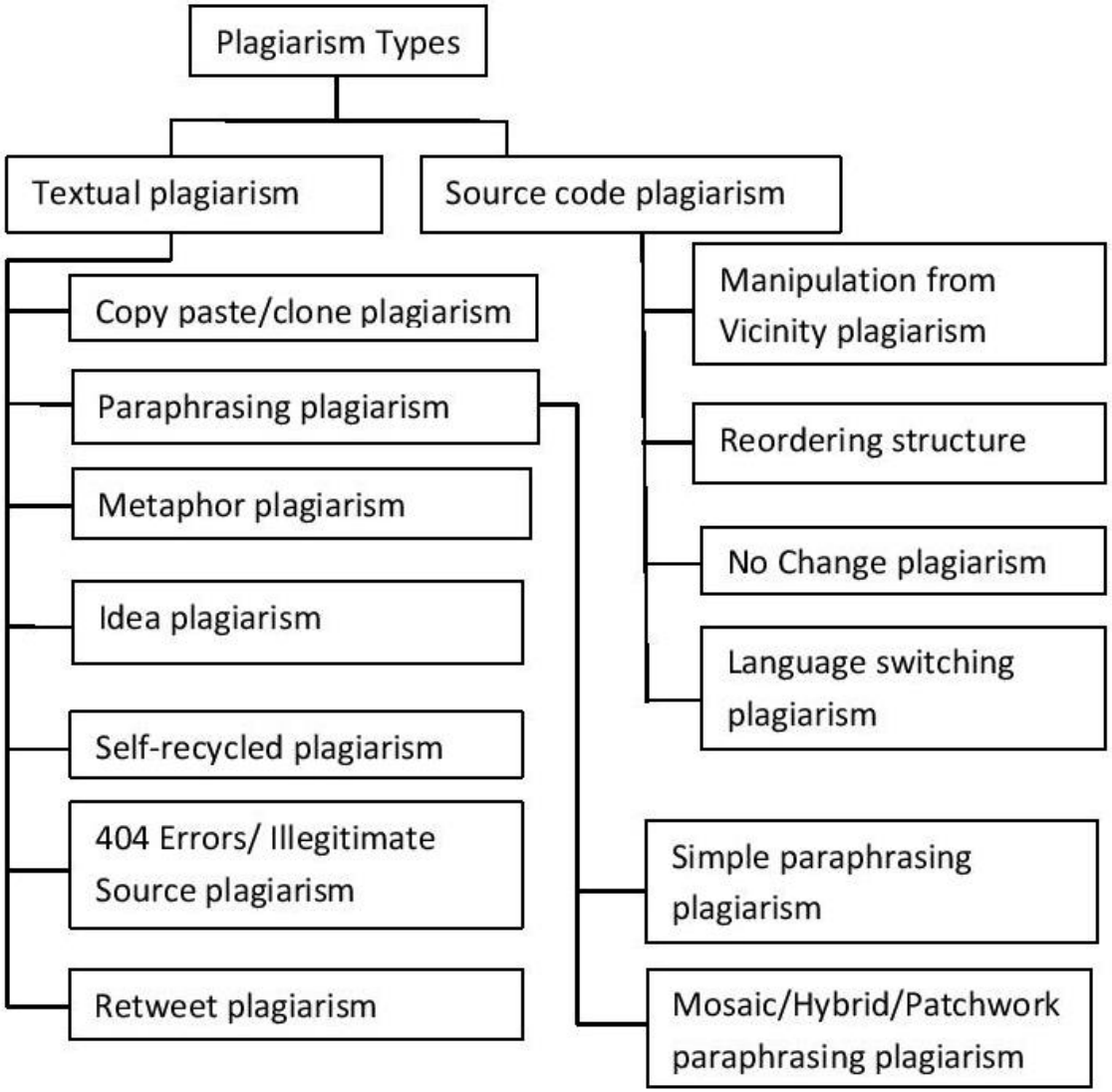
Data: First, Last
Result: Sum
while *(Last ≠ 0)* **do**
    Sum=First*Last;
    Last=Last-1;
end

Data: Start, Finish
Result: Total
while *(Finish ≠ 0)* **do**
    Total=Start*Finish;
    Finish=Finish-1;
end

(b)

Figure 1: Examples of (a) Textual (b) Source Code Plagiarism

Taxonomy of plagiarism



Different types of Textual Plagiarism I

Textual plagiarism further can be divided into seven sub categories based on its forms and application[8], [6]. We discuss each of these in brief, next.

1. *Deliberate copy-paste/clone plagiarism:* This type of textual plagiarism refers to copying other works and presenting as if your own work with or without acknowledging the original source.
2. *Paraphrasing plagiarism*: This form of plagiarism can occur in two ways as given below.

[6]C Barnbaum. Plagiarism: A student's guide to recognizing it and avoiding it.[online].[cit. 2010-12-14], 2009.
[8]Netra Charya, Kushagra Doshi, Smit Bawkar, and Radha Shankarmani. Intrinsic plagiarism detection in digital data 2(3), 23-30, 2015.

Different types of Textual Plagiarism II

i Simple paraphrasing: It refers to use of other idea, words or work, and presenting it in different ways by switching words, changing sentence construction and changing in grammar style.

ii Mosaic/Hybrid/patchwork paraphrasing: This form of textual plagiarism generally occurs when you combine multiple research contributions of some others and present it in a different way by changing structure and pattern of sentence, replacing words with synonyms and by applying a different grammar style without citing the source(s).

3. *Metaphor plagiarism:* Metaphors are used to present others idea in a clear and better manner.
4. *Idea plagiarism:* Here, idea or solution is borrowed from other source(s) and claiming as your own in a research paper.

Different types of Textual Plagiarism III

5. *Self/recycled plagiarism:* In this form, an author uses his/her own previous published work in a new research paper for publication.
6. *404 Error / Illegitimate Source plagiarism:* Here, an author cites some references but the sources are invalid.
7. *Retweet plagiarism:* In this form an author cites the reference of proper source but his/her presentation is very similar in the scene of original content wordings, sentence structures and/or grammar usage.

Different types of Source Code Plagiarism I

This type of plagiarism, as shown in previous figure can be divided into four subtypes.

1. *Manipulation from Vicinity plagiarism:* Here, a developer manipulates a program by (i) inserting, (ii) deleting, or (iii) substituting some codes in an existing program, with or without acknowledging the original source and claiming it as his/her own program.
2. *Reordering structure plagiarism:* In this type, the developer reorders the statements or functions of a program or changes syntax of a program without referring the original source.

Different types of Source Code Plagiarism II

3. *No change plagiarism:* Here, the developer adds or removes white spaces or comments or indentation of the program and claims the program as his/her own program.
4. *Language switching plagiarism:* In this type, the developer changes the languages, or a program written in one language is rewritten in another language and declare it as his/her own.

Based on characteristics, plagiarism can also be categorized into *literal* and *intelligent* plagiarism. Literal plagiarism consists of copy-paste/clone, paraphrasing, self/recycled, and retweet plagiarism. The other form of plagiarism can be considered as intelligent type of plagiarism.

Plagiarism Detection Approach I

Plagiarism can occur between two same or two different natural languages. Based on language homogeneity or heterogeneity of the textual documents being compared, the plagiarism detection can be divided into two basic types[3] i.e., monolingual and cross-lingual.

1. *Monolingual Plagiarism Detection:* This type of detection deals with homogeneous language settings e.g., English-English. Most detection methods are of this category. It can be further divided into two subtypes based on the use of external references during detection.

[3] 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 (Applications and Reviews), 42(2):133-149, 2012.

Plagiarism Detection Approach II

a Intrinsic Plagiarism Detection: This detection approach analyses the writing style or uniqueness of the author and attempts to detect plagiarism based on own-conformity or deviation between the text segments. It does not require any external sources for detection.

b Extrinsic Plagiarism Detection: Unlike the intrinsic approach, this approach compares a submitted research article against many other available relevant digital resources in repositories or in the Web for detection of plagiarism.

Plagiarism Detection Approach III

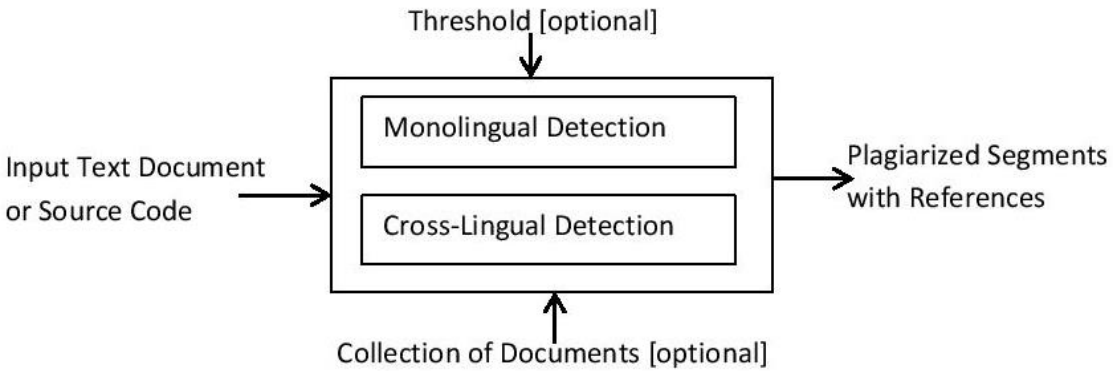
2. *Cross-Lingual Plagiarism Detection:* This detection approach is able to perform in heterogeneous language settings e.g., English-Chinese. There are only a few cross-lingual plagiarism detection methods available due to difficulty in finding proximity between two text segments for different languages.

Ways of Plagiarism Detection

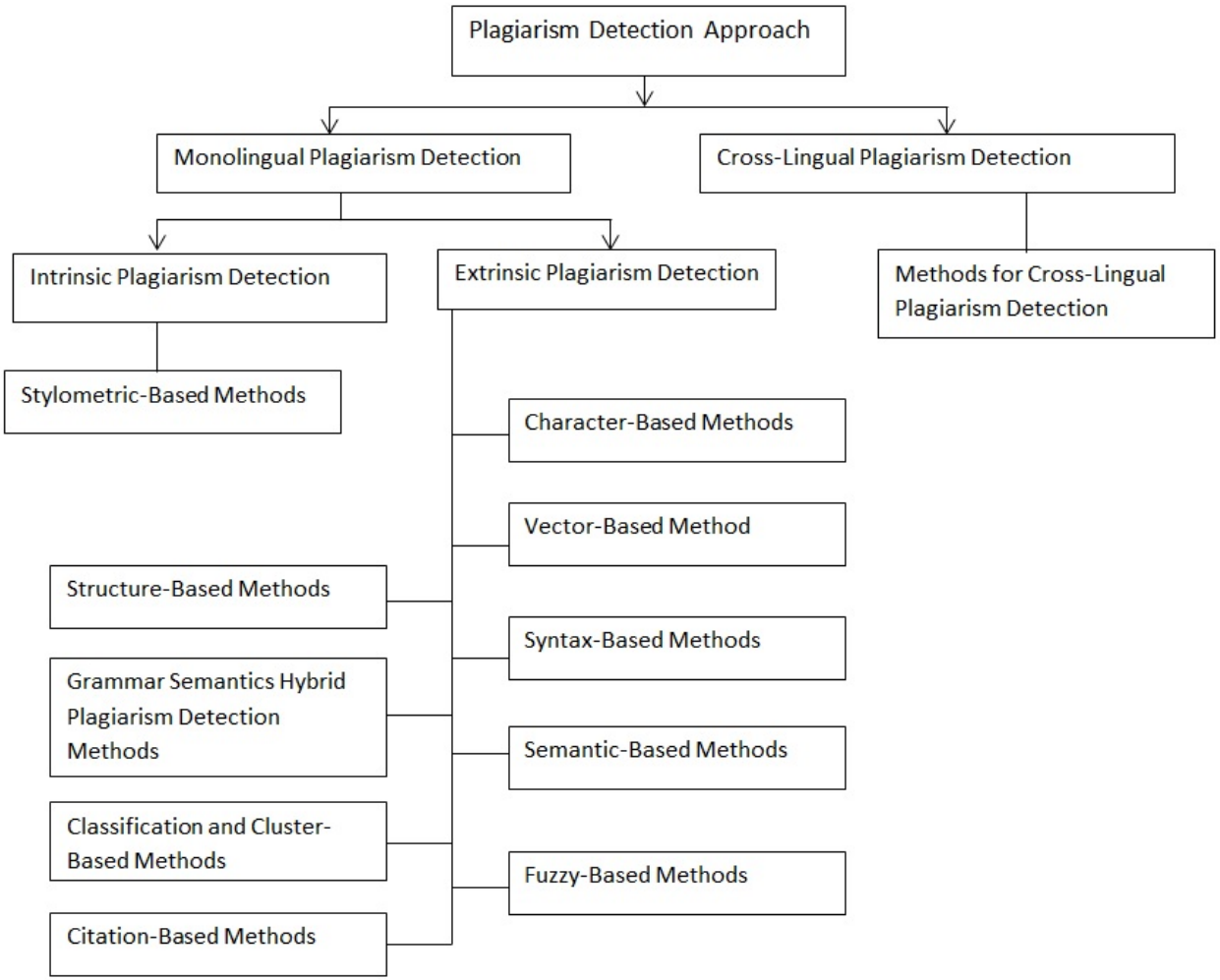
- There are different textual features like lexical feature, syntactic feature, semantic feature and structural feature, which can be used to detect similarity between two documents.
- Source code similarity detection can be carried out in various ways, such as (i) string matching, (ii) token matching, (iii) parse tree matching, (iv) program dependency graph (PDG) matching, (v) similarity-score matching and (vi) by hybridization of the above[1].

[1] Vale Alekya and S Sai Satyanarayana Reddy. Survey of programming plagiarism detection 2(6), 188-193, 2014.

A schematic view of a generic plagiarism detection method



Plagiarism Detection Methods: A Taxonomy



Character-Based Methods I

- It utilizes character-based, word-based, and syntax-based features to find similarity between a query document and an existing document.
- The similarity between a pair of documents may be estimated using both exact matching and approximate matching.
- Most of the detection techniques are developed based on n-gram or word n-gram based exact string similarity finding approach.

Character-Based Methods II

- Grozea et al. [15] use character 16-gram matching, whereas the authors of [7] use word 8-gram matching.
- One can use string similarity metric or vector similarity metric for finding similarity.

[7] Chiara Basile, Dario Benedetto, Emanuele Caglioti, Giampaolo Cristadoro, and MD Esposti. A plagiarism detection procedure in three steps: Selection, matches and âAsquaresâAI. In Proc. SEPLN, pages 19-23, 2009.

[15] Cristian Grozea, Christian Gehl, and Marius Popescu. Encoplot: Pairwise sequence matching in linear time applied to plagiarism detection. In 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse, page 10, 2009.

Vector-Based Method

- Here, lexical and syntax features are extracted and categorized as tokens rather than strings.
- The similarity can be computed using various vector similarity measures like Jaccard, Dice's, Overlap, Cosine, Euclidean and Manhattan coefficients.
- Cosine coefficient is useful in detecting partial plagiarism without sharing document content.
- Hence it is useful to detect plagiarism in documents where submission is considered as confidential[24].

[24] Haijun Zhang and Tommy WS Chow. A coarse-to-fine framework to efficiently thwart plagiarism. Pattern Recognition, 44(2):471-487, 2011.

Syntax-Based Methods I

- This methods exploit syntactical features like part of speech (POS) of phrase and words in different statements to detect plagiarism.
- The elements of basic POS tags are verbs, nouns, pronouns, adjectives, adverbs, prepositions, conjunctions, and interjections.

Syntax-Based Methods II

- In [11], [10], the authors use POS tag features followed by string similarity metric to analyze and calculate similarity between texts.
- Documents containing same POS tag features are carried out for further analysis and for identification of source of a plagiarism[10].

[10] Mohamed Elhadi and Amjad Al-Tobi. Use of text syntactical structures in detection of document duplicates. In Digital Information Management, 2008. ICDIM 2008. Third International Conference on, pages 520-525. IEEE, 2008.

[11] Mohamed Elhadi and Amjad Al-Tobi. Duplicate detection in documents and webpages using improved longest common subsequence and documents syntactical structures. In Computer Sciences and Convergence Information Technology, 2009. ICCIT'09. Fourth International Conference on, pages 679-684. IEEE, 2009. IEEE, 2008.

Semantic-Based Methods I

- Two sentences may be same but the order of their words may be different. For example, a sentence may be constructed by just transforming from active voice to passive voice.
- WordNet[12] is used in this context to find the semantic similarity between words or sentences.
- The degree of similarity between two words used in knowledge-based measures by Gelbukh[22] is calculated using information from a dictionary.

[12] Christiane Fellbaum. WordNet. Wiley Online Library, 1998.

[22] Sulema Torres and Alexander Gelbukh. Comparing similarity measures for original wsd lesk algorithm. Research in Computing Science, 43:155-166, 2009.

Semantic-Based Methods II

- Resnik[20] used WordNet to calculate the semantic similarity, whereas, Leacock's et al.,[17] determine semantic similarity by counting the number of nodes of shortest path between two concepts.

[17] Claudia Leacock, George A Miller, and Martin Chodorow. Using corpus statistics and wordnet relations for sense identification. Computational Linguistics, 24(1):147-165, 1998.

[20] Philip Resnik et al. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J. Artif. Intell. Res.(JAIR), 11:95-130, 1999.

Fuzzy-Based Methods

- Here, the words in a document are represented using a set of words of similar meaning and sets are considered as fuzzy since each word of the documents is associated with a degree of similarity[23].
- This method is attractive because it can detect similarity between documents with uncertainty.
- In [16], the degree of similarity between two documents or any two Web documents are identified by using fuzzy IR approach.

[12] Tommy WS Chow and MKM Rahman. Multilayer som with tree-structured data for efficient document retrieval and plagiarism detection. IEEE Transactions on Neural Networks, 20(9):1385-1402, 2009.

[19] MKM Rahman, Wang Pi Yang, Tommy WS Chow, and Sitao Wu. A flexible multi-layer self-organizing map for generic processing of tree-structured data. Pattern Recognition, 40(5):1406-1424, 2007.

Structure-Based Methods

- Unlike other methods discussed, structure based method uses contextual similarity.
- Contextual information is generally handled using tree-structure feature representation as can be found in ML-SOM [19].
- In [9], the authors detect plagiarism in two steps. First step performs document clustering and candidate retrieval using tree-structure feature representation and second step detects plagiarism by utilizing ML-SOM.

[9] Tommy WS Chow and MKM Rahman. Multilayer som with tree-structured data for efficient document retrieval and plagiarism detection. IEEE Transactions on Neural Networks, 20(9):1385-1402, 2009.

[19] MKM Rahman, Wang Pi Yang, Tommy WS Chow, and Sitao Wu. A flexible multi-layer self-organizing map for generic processing of tree-structured data. Pattern Recognition, 40(5):1406-1424, 2007.

Stylometric-Based Methods

- These methods aim to quantify the writing styles of the author to detect plagiarism.
- It computes similarity score between two sections or paragraphs or sentences based on stylometric features of the authors.
- These methods are instances of intrinsic plagiarism.
- The style representation formula may be writer specific or reader specific[27].
- One can find usefulness of outlier mining in this context to detect plagiarism in a document under this approach.

[27] Sven Meyer Zu Eissen, Benno Stein, and Marion Kulig. Plagiarism detection without reference collections. In Advances in data analysis, pages 359âĂŞ366. Springer, 2007.

Methods for Cross-Lingual Plagiarism Detection

- It requires in-depth knowledge of multiple languages.
- This type of methods work based on cross-lingual text features.
- Various types of these methods include (1) cross-lingual syntax-based methods, (2) cross-lingual dictionary-based methods, and (3) cross-lingual semantic-based methods [3].
- In [15], a statistical model is used to evaluate the similarity between two documents regardless of the order in which the terms appear in suspected and original documents[18].

[18] Ahmed Hamza Osman, Naomie Salim, and Albaraa Abuobieda. Survey of text plagiarism detection. Computer Engineering and Applications Journal (ComEngApp), 1(1):37-45, 2012.

Grammar Semantics Hybrid Plagiarism Detection Methods

- These methods eliminate the limitations of semantic-based methods.
- They are capable of detecting copy/paste and paraphrasing plagiarism accurately.
- A semantic-based method usually cannot detect and determine the location of plagiarised part of the document but such grammar-based method can address this issue efficiently [5], [3].

[3] 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 (Applications and Reviews), 42(2):133-149, 2012.

[5] Jun-Peng Bao, Jun-Yi Shen, Xiao-Dong Liu, and Qin-Bao Song. A survey on natural language text copy detection. Journal of software, 14(10):1753-1760, 2003.

Classification and Cluster-Based Methods

- In information retrieval process, supervised and unsupervised grouping of documents play an important role.
- It helps in reducing the document comparison time significantly during plagiarism detection[26].
- Some methods [25], [21] use keywords or specific words to cluster the similar sections of documents.

[21] Antonio Si, Hong Va Leong, and Rynson WH Lau. Check: a document plagiarism detection system. In Proceedings of the 1997 ACM symposium on Applied computing, pages 70-77. ACM, 1997.

[25] Manuel Zini, Marco Fabbri, Massimo Moneglia, and Alessandro Panunzi. Plagiarism detection through multilevel text comparison. In 2006 Second International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution (AXMEDIS'06), pages 181-185. IEEE, 2006.

Citation-Based Methods

- It detects plagiarism in documents based on citations.
- In [13], a novel method is proposed to detect plagiarism on the basis of citation.
- Citation-based methods belong to semantic plagiarism detection techniques because these techniques use semantics contained in the citation in a document[14].
- The similarity between two document is computed based on the similar patterns in the citation sequences[14].

[13] Bela Gipp and Joran Beel. Citation based plagiarism detection: a new approach to identify plagiarized work language independently. In Proceedings of the 21st ACM conference on Hypertext and hypermedia, pages 273-274. ACM, 2010.

[14] Bela Gipp and Norman Meuschke. Citation pattern matching algorithms for citation-based plagiarism detection: greedy citation tiling, citation chunking and longest common citation sequence. In Proceedings of the 11th ACM symposium on Document engineering, pages 249-258. ACM, 2011.

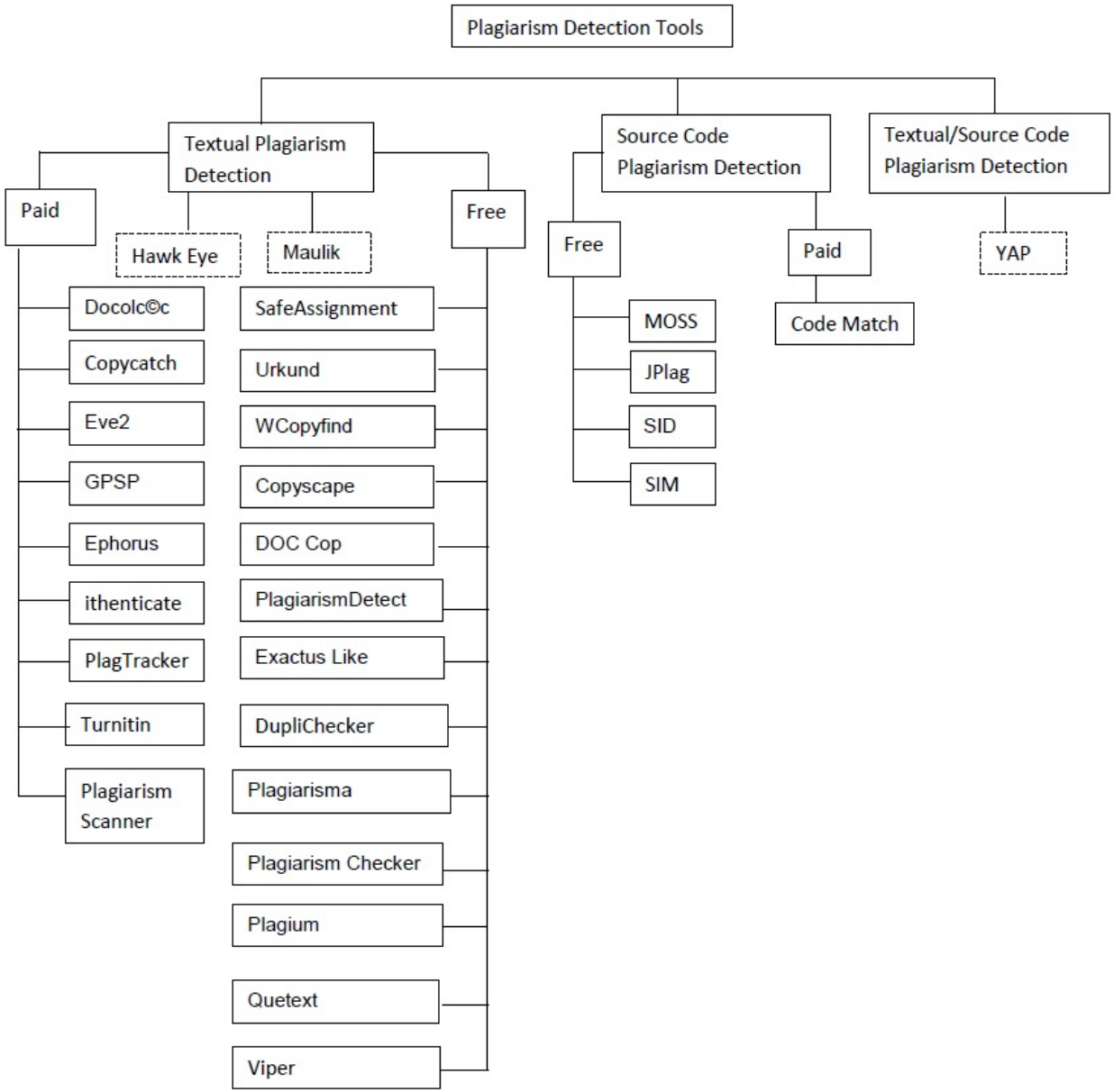
Plagiarism Detection Techniques: A General Comparison

Table 1: PLAGIARISM DETECTION TECHNIQUES: A General Comparison

Author & Name	Intrinsic(I)/Extrinsic	Approach used	model used		Language(s)	Types of plagiarism								References
			Mono-Lingual IR	Cross-Lingual IR		Literal			Intelligent					
						Copy	Near copy	Restructuring	Paraphrasing	Summarising	Translating	Idea(Section)	Idea(Context)	
Character-Based (CNG)	E	String Matching	✓		any	✓	✓							
Vector-Based(VEC)	E	Text Similarity	✓		any	✓	✓	✓						
Syntex-Based(SYN)	E	Text Similarity	✓		specific	✓	✓	✓						
Semmantic-Based(SEM)	E	Word Similarity and Local Semantic Density	✓		specific	✓	✓	✓	✓	?				
Fuzzy-Based(FUZZY)	E	Fuzzy set of synnony words	✓		specific	✓	✓	✓	✓	?				
Structural-Based(STRUC)	E	Tree-Structured Features Representation	✓		specific	✓	✓	✓	?	?		?	?	
Stylometric-Bsed(STYLE)	I	Author vocabulary richness and style complexity	✓		specific	✓	✓	✓						
Cross-Lingual(CROSS)	E	Cross-Lingual Syntax, semantic, dictionary, statistic		✓	cross						✓			
Grammar-Based(GRAM)	E	String Matching	✓		any	✓	✓							
Cluster-Based(CLUS)	E	Text summerization and exact matching			specific	✓	✓	✓						
Citation-Based(CITE)	E	Word Similarity and Local Semantic Density	✓		specific	✓	✓	✓	✓	?				

Note: IR: Information Retrieval, ?: Need further research

Plagiarism Detection Tools Classification



Plagiarism Detection Tools: A General Comparison

Table 2: PLAGIARISM DETECTION TOOLS: A General Comparison

Name	Year	P1	P2	P3	P4	Source
SafeAssignment	2008	E	Y	S	N	http://www.safeassign.com/
Docol©	2005	E	Y	S	N	https://www.docoloc.de/
Urkund	2000	E	Y	S	N	http://www.urkund.com/
Copycatch	2002	I/E	Y	S	N	www.copycatchgold.com
Wcopyfind	2004	I/E	Y	S	N	http://www.plagiarism.phys.virginia/
EVE2	*2001*	E	Y	S	Y	www.canexus.com
GPSP	1999	I	Y	S	Y	http://www.plagiarism.com/
MOSS	1994	E	N	M	N	http://theory.stanford.edu/~aiken/moss/
Jplag	1997	E	Y	M	Y	https://jplag.ipd.kit.edu/
Copyscape	2011	E	Y	S	Y	http://www.copyscape.com/
DOC Cop	2006	E	--	S	N	www.doccop.com/
Ephorus	---	E	Y	S	N	http://www.ephorus.com/
iThenticate	1996	E	Y	S	N	http://www.ithenticate.com/
PlagiarismDetect	2008	E	Y	S	N	plagiarismdetect.org/
Exactus Like	2016	E	Y	S	N	http://like.exactus.ru/index.php/en
DupliChecker	2006	E	Y	S	N	www.duplichecker.com/
Plagiarisma	---	E	Y	S	N	http://plagiarisma.net/
PlagiarismChecker	2006	E	Y	S	N	http://www.plagiarismchecker.com/
Plagium	2006	E	Y	S	N	http://www.plagium.com/
PlagTracker.	2011	E	Y	S	N	http://www.plagtracker.com/
Quetext	---	I/E	Y	S	N	http://www.quetext.com/
Turnitin	2000	E	Y	S	N	http://www.turnitin.com/
Viper	2007	E	Y	S	Y	http://www.scanmyessay.com/
Maulik	2016	E	?	S	N	---
Plagiarism Scanner	2008	E	Y	S	N	http://www.plagiarismscanner.com/
Hawk Eye	2015	E	Y	S	--	---
Code Match	---	--	Y		Y	http://www.safe-corp.com/
SID	2004	E	Y	S	Y	http://software.bioinformatics.uwaterloo.ca/SID/.
SIM	1999	E	N	M	Y	http://www.cs.vu.nl/dick/sim.html
YAP3	1996	E	?	M	N	---
PlagScan	2015	E	Y	S	N	www.plagscan.com/

P3: Intrinsic /Extrinsic, P2: User friendly? P3: Submission of single/multiple files?
P4: Source code availability?
E: Extrinsic, I: Intrinsic, Y: Yes, N: No, S: Single, M: Multiple

Issues and Challenges I

- Most prominent methods have been able to address the major issues related to (i) salient syntactic and semantic feature extraction, (ii) handling of both monolingual and cross-lingual plagiarism detection, and (iii) detecting plagiarism in both text data and program source code with or without using references.
- Due to the rapid growth of digital technology to support its reproduction, storage and dissemination, some important issues and research challenges are still left unattended.
- We highlight some issues and challenges that need to be addressed by computer science and linguistic researchers.

Issues and Challenges II

(a) A detection method for both text data and source code that ensures both proof of correctness and proof of completeness is still missing, and hence an important issue.

(b) A proximity measure that guarantees detection of plagiarized text segment(s) in both intrinsic and extrinsic detection framework with high accuracy, is still not available.

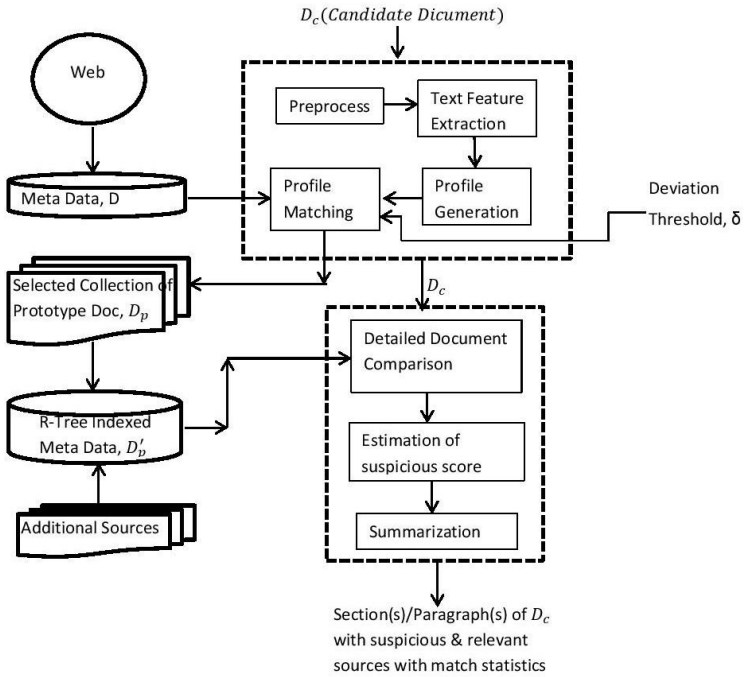
(c) Developing a cross-lingual plagiarism checking tool that can perform without external references but ensures high accuracy is a challenging task.

(d) Developing a repository that maintains references based on author footprints, which is complete and accurate is another challenging task.

Issues and Challenges III

(e) Developing a plagiarism checker that accepts an idea narrated by user and generates a detail plagiarism report (with similarity if detected from 1%-99%) with correct sources, is an important issue.

Conceptual Framework for Plagiarism Detection I



Conceptual Framework for Plagiarism Detection II

- This conceptual framework is a hybrid approach to detect text plagiarism with high accuracy at an early stage.
- The approach is designed based on the merit of both intrinsic and extrinsic plagiarism detection approaches.
- Initially it takes a candidate document D_c as input, computes its salient global features, segments D_c and extract local features for each segmented text, and finally forms a k-dimensional ($k = m(global) + n(local)$) feature vector ${D_c}^i$.

Conceptual Framework for Plagiarism Detection III

- The global feature values of the feature vector are used to identify the a subset of relevant source documents for a given input candidate document, whereas, the local feature values are utilized during matching to identify the possible plagiarised text segments as well as the relevant source documents.
- Our approach assume the possession of such k dimensional feature vector against each prototype document ${D_p}^j$ in the repository.
- The following features make our approach more attractive while comparing its other counterparts.
  a. It exploits the merits of both intrinsic and extrinsic plagiarism in a cost effective manner.

Conceptual Framework for Plagiarism Detection IV

b The nearest-neighbour based outlier mining technique is able to detect a plagiarised text segment or paragraph at an early stage in terms of non-conformity.

c The meta-level (i) initial reference filtering and (ii) the final selection of sources or subset of references relevant to the plagiarised text segment, makes our approach more economic and computationally cost-effective.

d The flexibility of allowing one to incorporate or delete features while constructing vector to represent a text document in general and the uniqueness of a text segment.

Conclusions I

- We reported an exhaustive survey on plagiarism detection methods and tools in a systematic way.
- We presented a taxonomy of various forms of plagiarism occur in text data and source code.
- Next, we reported a large number of methods and tools under various categories and compared and analyzed their pros and cons.
- We present a hybrid plagiarism detection approach based on the merit of both intrinsic and extrinsic plagiarism detection approaches.

Conclusions II

- Finally, we have highlighted a list of issues and research challenges towards developing a plagiarism checker that is complete and correct for both monolingual and cross-lingual text data and for source code.

References I

[1] Vale Alekya and S Sai Satyanarayana Reddy. Survey of programming plagiarism detection, 2(6), 188–193, 2014.

[2] Asim M El Tahir Ali, Hussam M Dahwa Abdulla, and Vaclav Snasel. Overview and comparison of plagiarism detection tools. In *DATESO*, pages 161–172. Citeseer, 2011.

[3] 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 (Applications and Reviews)*, 42(2):133–149, 2012.

[4] Melissa S Anderson and Nicholas H Steneck. The problem of plagiarism. In *Urologic Oncology: Seminars and Original Investigations*, volume 29, pages 90–94. Elsevier, 2011.

References II

[5] Jun-Peng Bao, Jun-Yi Shen, Xiao-Dong Liu, and Qin-Bao Song. A survey on natural language text copy detection. *Journal of software*, 14(10):1753–1760, 2003.

[6] C Barnbaum. Plagiarism: A student's guide to recognizing it and avoiding it.[online].[cit. 2010-12-14], 2009.

[7] Chiara Basile, Dario Benedetto, Emanuele Caglioti, Giampaolo Cristadoro, and MD Esposti. A plagiarism detection procedure in three steps: Selection, matches and âĂIJsquaresâĂİ. In *Proc. SEPLN*, pages 19–23, 2009.

[8] Netra Charya, Kushagra Doshi, Smit Bawkar, and Radha Shankarmani. Intrinsic plagiarism detection in digital data, 2(3), 23–30, 2015.

References III

[9] Tommy WS Chow and MKM Rahman. Multilayer som with tree-structured data for efficient document retrieval and plagiarism detection. *IEEE Transactions on Neural Networks*, 20(9):1385–1402, 2009.

[10] Mohamed Elhadi and Amjad Al-Tobi. Use of text syntactical structures in detection of document duplicates. In *Digital Information Management, 2008. ICDIM 2008. Third International Conference on*, pages 520–525. IEEE, 2008.

[11] Mohamed Elhadi and Amjad Al-Tobi. Duplicate detection in documents and webpages using improved longest common subsequence and documents syntactical structures. In *Computer Sciences and Convergence Information Technology, 2009. ICCIT'09. Fourth International Conference on*, pages 679–684. IEEE, 2009.

[12] Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.

References IV

[13] Bela Gipp and Jöran Beel. Citation based plagiarism detection: a new approach to identify plagiarized work language independently. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, pages 273–274. ACM, 2010.

[14] Bela Gipp and Norman Meuschke. Citation pattern matching algorithms for citation-based plagiarism detection: greedy citation tiling, citation chunking and longest common citation sequence. In *Proceedings of the 11th ACM symposium on Document engineering*, pages 249–258. ACM, 2011.

[15] Cristian Grozea, Christian Gehl, and Marius Popescu. Encoplot: Pairwise sequence matching in linear time applied to plagiarism detection. In *3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse*, page 10, 2009.

References V

[16] Jonathan Koberstein and Yiu-Kai Ng. Using word clusters to detect similar web documents. In *International Conference on Knowledge Science, Engineering and Management*, pages 215–228. Springer, 2006.

[17] Claudia Leacock, George A Miller, and Martin Chodorow. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics*, 24(1):147–165, 1998.

[18] Ahmed Hamza Osman, Naomie Salim, and Albaraa Abuobieda. Survey of text plagiarism detection. *Computer Engineering and Applications Journal (ComEngApp)*, 1(1):37–45, 2012.

[19] MKM Rahman, Wang Pi Yang, Tommy WS Chow, and Sitao Wu. A flexible multi-layer self-organizing map for generic processing of tree-structured data. *Pattern Recognition*, 40(5):1406–1424, 2007.

References VI

[20] Philip Resnik et al. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.(JAIR)*, 11:95–130, 1999.

[21] Antonio Si, Hong Va Leong, and Rynson WH Lau. Check: a document plagiarism detection system. In *Proceedings of the 1997 ACM symposium on Applied computing*, pages 70–77. ACM, 1997.

[22] Sulema Torres and Alexander Gelbukh. Comparing similarity measures for original wsd lesk algorithm. *Research in Computing Science*, 43:155–166, 2009.

[23] Rajiv Yerra and Yiu-Kai Ng. A sentence-based copy detection approach for web documents. In *International Conference on Fuzzy Systems and Knowledge Discovery*, pages 557–570. Springer, 2005.

[24] Haijun Zhang and Tommy WS Chow. A coarse-to-fine framework to efficiently thwart plagiarism. *Pattern Recognition*, 44(2):471–487, 2011.

References VII

[25] Manuel Zini, Marco Fabbri, Massimo Moneglia, and Alessandro Panunzi. Plagiarism detection through multilevel text comparison. In *2006 Second International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution (AXMEDIS'06)*, pages 181–185. IEEE, 2006.

[26] Du Zou, Wei-Jiang Long, and Zhang Ling. A cluster-based plagiarism detection method. In *Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September*, 2010.

[27] Sven Meyer Zu Eissen, Benno Stein, and Marion Kulig. Plagiarism detection without reference collections. In *Advances in data analysis*, pages 359–366. Springer, 2007.

Thank You...