Iranian Journal of Military Medicine Vol. 14, No. 3, Autumn 2012; 168-177

Plagiarism: Concepts, Factors and Solutions

Bahadori M.[1] *PhD*, Izadi M[1] *MD*, Hoseinpourfard M[1*] *PhD*

[1] Health Management Research Center, Baqiyatallah University of Medical Science, Tehran, Iran

Abstract

The goal of knowledge production is the discovery of facts and improving the human situation, and as such, plagiarism and using other unethical means are not compatible with this goal. Most academic scholars agree that plagiarism is a serious violation of publishing ethics. In recent decades, the scientific community has become really concerned about the fast growth of plagiarism. Although plagiarism is widespread, it isn't consistent with the principles of science.

Nowadays some media publish worrying news of plagiarism in scientific publications, including data manipulation by well-known scientists. The prevalence rate of plagiarism has been reported in different studies turns out to be different in various fields, countries, educational levels and times.

The goal of this study is to review the scientific concepts related to plagiarism, its factors and roots, its prevalence in the world and methods of detecting it in order to improve the awareness of instructors and students of plagiarism.

Keywords: Plagiarism, Solutions, Concepts, Causes

Introduction

The goal of research is to produce knowledge, and the aim of producing knowledge is to improve human situation while doing research using unethical or inappropriate means leads to scientific corruption, which is against scientific knowledge production [1]. No doubt, there is plagiarism in the scientific community although it is against basic scientific principles. Plagiarism is useless, meaningless, unethical and thus forbidden [2].one of the pathological componenets in the relationships between people is the legal culture in society [3].

Unethical issues are quickly increasing in the realm of science. In the future, such issues in gathering data, cooperation between scientists and in publications will most probably get more complicated and more difficult to deal with. More than ever before, postgraduate medical students should know about methods, technologies and concepts of science. The global competition among the scientists of developing countries, especially Asian ones, is a new reality for the western researchers who want to be the best in all areas of research. Researchers in developing countries are increasingly enjoying more research budgets, and this development has been accompanied by governmental and institutional demand for better results and more publications in scientifically accredited journals [4].

Plagiarism is a controversial issue in higher education, and it is increasingly widespread among students. Some challenges in academic activities are due to the increase in the number of students [5]. Today, open access publications are not only reasonable but also very vital to scientific innovations. Unlimited access to scientific ideas, methods, findings and results is not compatible with the restricting regulations of copyright, and this has made for more plagiarism [6].

"If plagiarism turns into an ordinary and usual activity, it will affect the security of scientific knowledge and destroy all social realms. In such a situation, nobody will bother doing research; rather, everybody will make use of ready-made knowledge produced by the past researchers and will destroy all knowledge. Such unreasonable behavior will devastate the foundations of scientific progress and everything else. And if a country loses its firm scientific foundations, it will remain in past achievements and will not experience progress" [7].

Plagiarism is one of the important issues of universities in recent years. In the last two decades, the

[*]**Corresponding author**: Hoseinpourfard M.J., Please, direct all correspondence at hpf.javad@gmail.com.

progress in computer technology, that is, running websites to provide university services, the copy-paste tool, and loads of pre-fabricated papers, has made for an increase in plagiarism [8]. Nowadays some media publish worrying news of plagiarism in scientific publications, including data manipulation by well-known scientists. The ethics of scientific publication is in direct connection to the concepts of copyright in writing scientific papers and of plagiarism. Sometimes, journal editors take the writers' cunningness for their lack of familiarity with journal regulations or their lack of attention to a certain paper. As Kosovsky notes, "the road to hell is paved with good intentions" and after that, the writers make very serious ethical mistakes to the end [9].

The author of a book, paper, poem or a scientific passage, after hours of thinking and writing about a subject, puts to paper the fruit of years of his or her continuous efforts. As such, the plagiarist not only steals the fruit of such efforts but also registers all that painstaking work to his or her own name [7].

Plagiarism is hundreds of years old, but, due to the progress in information technology, it has acquired new and different methods compared to the past. Plagiarism was almost a rare phenomenon until 1990, but it has spread across the world in recent years and has worried the academic community [10]. In the past, there were a few scientists who produced knowledge and some of them would produce no more than a couple of papers in their lifetime. In those times, strict reviewing principles were at work, there just a few journals and scientists had a hard time convincing the scientific community to accept their ideas. In the 19th century, the problem was stealing ideas, and that was why many discoveries and inventions were disputed. Today, however, the number of scientists, students, journals and papers has really increased. While there is no problem with the increase in the number of papers, peer-reviewing of the papers is the main problem. It is certainly expected of a reviewer to have a good command of the subject of a paper. But, given the large number of papers to be reviewed, are there enough specialists to review the papers? No scientist can claim that he or she has studied all specialist papers in his or her area of knowledge, and this paves the way for some plagiarists to take advantage of the situation [11].

Ben Jonson was the first one to sue the term plagiarism in the early 17th century. It was hard for authors to protect their writings before devising copyright laws. But as plagiarism increased in the 18th century and copyright laws were consequently clearly defined and devised by the middle of the century, plagiarists faced a change in the public opinion and strong ethical viewpoints towards plagiarism [12].

In view of the prevalence of plagiarism in the scientific community and its devastating effects on scientific progress, this study aims at surveying the concepts, causes and solutions to the issue of plagiarism.

Terminology, Definitions and Idioms

According to the Persian dictionary of Dehkhoda, the word "steal" means "taking away somebody's possession with deception and tricks" or "to take hold of something without the right to do so" [13]. Wilson Mizner states that "when we steal an idea from one author, it will be called plagiarism, but when we do it from a few authors, it is called research" [14].The word plagiarism comes from the word "plagarius", meaning kidnapper, robber, misleader, and literary thief" [15]. Plagiarism usually refers to stealing ideas or words that are higher than the level of public knowledge [16].

In Webster's Dictionary, a plagiarist is defined as "One who plagiarizes, or purloins the words, writings, or ideas of another, and passes them off as his own; a literary thief" [17], and plagiarism as "taking someone›s words or ideas as if they were your own"[18]. The University of Liverpool defines plagiarism as the "use of materials from unacknowledged sources or direct quotation of materials from documented references without acknowledging that the words have been taken verbatim from those references" [19]. Payer sees plagiarism as "taking others' ideas, words or wok as if they were your own" [20]. Or as Stebel man puts it, plagiarism consists of "claiming as your own the writings and research papers that originally belong to others [21]. Vessal and Habibzadeh take plagiarism to be "ascribing others' ideas, processes, results and words to oneself without due acknowledgement" [22]. Using sentences from published medical literature with little change in the words without acknowledging the source is also an in-

stance of plagiarism. Using unpublished images or pictures with the owners' permission is also called plagiarism [23].

The Federal Government of the United States defines "research misconduct as fabrication, falsification or plagiarism in proposing, implementing or reviewing of research projects or in reporting the results of research" [24].

Plagiarism is an unethical activity in scientific writing. For something to be called plagiarism, it needs to be a serious deviation from normally accepted behavior of the relevant scientific community which is done consciously and deliberately and must be proved with solid evidence. Plagiarism may occur in different forms: stealing ideas and stealing parts of texts. Self-plagiarism happens when an author uses his or her own previously published work without acknowledging it [25]. Self-plagiarism is defined in three ways in the relevant literature: 1) publishing a paper which basically overlaps another paper without due acknowledgement; 2) breaking a large paper into a few smaller papers and publishing them separately, called salami slicing and 3) republishing the same work. Copyright, on the other hand, means enhancing knowledge and useful arts by providing limited-time security for authors and inventions through exclusive rights regarding their writings and inventions. Authors of technical papers are usually asked to transfer the copyright of their work to the journal or the publisher [26].

Scientific integrity depends on honesty and transparency of the methods of producing and transferring knowledge [26]. Republishing results is announcing the same results in two or more papers, multiple recalculations of the same results in meta-analyses and as a result in serious errors in research [27].

Duplicate or redundant publication occurs when there is an overlap, without acknowledging it, between two papers in terms of their hypotheses, data, arguments or results. This could include an overlap with other authors, their results or their samples. The most important cases involve lack of acknowledging the sources. The following are example cases of republishing: publishing data which has been published before, reusing tables and figures in later publications, publishing larger papers using previous smaller papers, publishing the same data in two papers (on with a clinical focus and one with a theoretical focus), and publishing the same paper under two names, one being the real author in his or her own country and the other being a foreign author.

Republishing, which is done in a deceptive way, is certainly unacceptable. If editors, reviewers and end readers of data notice the overlap between papers, they can make the right decision about it. Duplicate publication is, nevertheless, deceptive and involves three problems: it is unethical, it wastes resources and it has adverse impacts upon future clinical and research decisions. Editors and readers of a published report want to make sure that they are dealing with new and important data, and may wrongly be persuaded to think so, while this is not the case. Duplicate or redundant publication misleads the readers and reduces the credibility of the journal as well as its ability to attract good papers. Duplicate publication makes for wasting resources by wasting the time which should be allocated to other papers [28].

"Most academic researchers agree that plagiarism is a serious problem in the ethics of publication. Plagiarism appears in different forms: stealing ideas and stealing texts (verbatim plagiarism). Plagiarism is no doubt an instance of misconduct. Stealing part of text and rephrasing it is a severe problem in the humanities and literature where innovation in phrasing and eloquence are essential. But in the realm of science, it is the scientific content itself, not its eloquence, that matters" [29].

The purpose of scientific journals is to some extent different from that of non-scientific ones. For instance, medical journals are published in order to improve the science of medicine and public health by publishing the results of scientific research. In many areas such as literature and humanities, however, different authors have different views. They try to reflect their own understanding and feelings of texts by means of a selection of good and suitable words. Thus, each and every word, along with its immediate context, has a role in conveying the meaning to the reader. But in a scientific writing, the writer's audience consists of scholars who are looking for facts based on solid evidence. Therefore, the writer is supposed to observe and report correctly. Unlike literary researchers, a scientific paper author should follow a certain and well-es-

tablished scientific method and make sure that he or she will not be become biased in his or judgments since this can endanger the truth or reliability of the judgments. Thus, whether or not he or she is eloquent, as far as an author is a just observer who works based on accepted scientific methods, evidence and facts, he or she can publish his or her findings and could be said to have followed a universally accepted method [29].

Plagiarism, in general, includes attributing somebody else's work to yourself without giving credit to the author, copying other's ideas or words without giving credit to the source, not putting quotations in quotation marks, giving the wrong information about a reference, changing the words while keeping the structure of a sentence from another source without acknowledging it, and copying a large number of words or ideas from other sources with or without due acknowledgement [30].

Another definition of plagiarism numerates the ways of plagiarizing in the following way: "'copy-past' which means verbatim copying of words, plagiarizing ideas, which consists of using a concept or idea which is not commonly known to others, rephrasing, which means changing the grammatical structure, using synonyms, reordering the original sentences, or rewriting the same content in different words, artistic plagiarism, which denotes presenting others' works using a different medium such as text, voice, or image, plagiarizing codes, that is, using other programs' codes, algorithms and functions without the right permission or referencing, using expired or neglected links, adding quotation marks or other referencing signs without providing the right referencing information or updating links to sources, inappropriate use of quotation marks, failure to recognize the quoted parts of a text, incorrect referencing, i.e., adding incorrect referencing information or references which do not exist and plagiarism in translation, which consists of translating a text without giving reference to the original text" [30].

The following are some instance of student plagiarism: stealing material from a source and passing it for as their own, for instance, by buying a preordered paper, copying an entire paper without acknowledging it, presenting another student's work without their knowledge, presenting somebody else's paper and passing it as your own, copying materials of one or more texts and providing the right citations without using quotation marks to make the readers believe that they have paraphrased the materials not quoted them, and rewording sentences from other sources without giving credit to them [31].

Recognizing plagiarism faces a number of problems. One problem is recognizing the amount of plagiarism because it can cover a wide scope. The second problem is the question as to how much change in the original material can make for plagiarism [31]. Roig argues that many students struggle between rewording and summarizing because they cannot distinguish between them. The third issue is that most authors believe that there is no need to reference common knowledge, but we may ask what common knowledge is and who defines it? [32]

Plagiarism can be divided into two types with regard to intentions. The first type is intentional plagiarism where the author is fully aware of the plagiarism and is willing to do it. The second type of is unintentional plagiarism where a person plagiarizes due to his or her unawareness and lack of skill in writing. The latter type could be prevented [33].

In another classification, plagiarism is divided into four categories: 1) "casual plagiarism, which occurs because of lack of awareness of plagiarism, or insufficient understanding of referencing or citation;" 2) unintentional plagiarism, where, due to the wide amount of knowledge in the scientific area, a person may unknowingly present ideas similar to those of others;" 3) intentional plagiarism, where a person deliberately and knowingly copies part or all of somebody else's work without giving credit to them; and 4) self-plagiarism, which consists of reusing one's own published work in a different form with acknowledging it" [30].

The Prevalence of Plagiarism

Researches show that plagiarism is an increasingly widespread practice in educational and research institutes. The rate of plagiarism is different in various areas of research. As reported, the rates of prevalence of plagiarism are 78% in the students of Organizational Studies and 63% in the students of humanities. Also, there is a meaningful difference between the behavior of American students and that of Hungary in terms of plagiarism [31].

Similarly, studies carried out by Park in the United States, South Africa and Finland reveal that the rates of plagiarism are different for different areas of study [31]. According to some research, the number of plagiarizing students in an institute increased from 11% in 1963 to 49% in 1993. These results include all forms of plagiarism, including copying material from encyclopedias, journals, papers and the like [34]. Jude Carroll argues that unacknowledged copying of materials from books and journals are more common than from web sites [35].

According to some research, 12% of the papers suspected of plagiarism belong to the students of Politics. According to another researcher, in an American university, 16.5% students report to have plagiarized, and 50% of the students believed that their classmates often copy-pasted material from the Internet without acknowledging it [36]. Satterthwaite argues that the rate of plagiarism in America is 30% [37]. One study shows that 94% of students had misconducted in their research for at least once, and another study shows this rate to be 91% [38]. Dordoy, who has studied plagiarism in the students of an English university, claims that the rate of copying a paragraph from a book or a web site was 73.9% [39].

A study focusing on plagiarism reveals that 48% of the students were not aware of the methods and requirements of referencing [40]. The results of some research on academic misconduct tell us that 76% of students had responded positively to cheating in high school or college [40]. Carroll holds that because most students do not know what makes for plagiarism, they do not commit it with the intention of deceiving others [41]. A study in 2009 indicates that 212 papers showed some potential signs of plagiarism. In these papers, the similarity between the original paper and the republished one was 86.2% while the average of shared sources was 73.1%. Of the 212 papers, only 47 (22%) cited the original paper. Also, there were miscalculations, contradictory data and manipulation of figures in 47% of the papers [42].

Bloemenkamp et al. report that 20% of the papers published in Holland's Journal of General Medicine had already been published elsewhere. Similarly, Schein and Paladugu reported that one sixth of the papers appearing in three journals of operation showed some signs of duplicate publication. According to Tramer et al, 17% of the reports and 28% of patient data were duplicated and the inclusion of redundant data in a meta-analysis led to a 23% overestimation of the treatment effectiveness of an antiemetic agent. Redundant publication can undermine the results of studies which are based on reliable evidence. It can exaggerate the significance of the results and mislead the reader [28]. According to a met-analysis by Fanelli, medical scholars report more cases of scientific misconduct that those of fields of study [23]. The University of Sao Paolo has appeared in the media on the suspicion of plagiarism in scientific publication and research. Journals are concerned about fabrication or making up of data in published papers or duplication production of data or text by other authors without proper citation or referencing or even duplication of the published research or texts in other papers [43].

Factors of Plagiarism

According to Ashworth, the concept of plagiarism is not clear enough so much so that some students are afraid of unwitting plagiarism while putting to paper what they take to be their own ideas [44]. Researches show that students and teachers have different understandings of plagiarism. For some teachers, some definitions are influenced by higher education values such as the copyright, personal effort and unity in the university [45]. The multiplication of databases, with all its benefits, has also caused a rapid growth in plagiarism. Some factors affecting student attitudes toward plagiarism are ignorance, lack of personal investment in their education, situational ethics, and lack of consistent styles among and within various disciplines [46]. According to Dordoy, the most important factors influencing plagiarism include promotion, laziness or mismanagement of time, easy access to materials on the Internet, unawareness of rules and regulations and unwitting plagiarizing [39].

Some other factors causing plagiarism are low commitment to the learning process and focusing on getting an academic degree, the student life style, family pressures, etc. make students try to achieve the best results with the least efforts and in the least time [47]. In the past, students had to go to libraries, retrieve information and retype it while today and with the rapid progress of the Internet,

this process has changed and most teachers believe that computers have made it easier to cheat and plagiarize [48].Angellil-Carter claims that there is no transparency about factors influencing plagiarism all over a university [49]. Dickert claims that not only are Hong Kong university students not familiar with plagiarism but also it is very hard to detect plagiarism in this university [50]. Information is easily accessible through electronic media and word processing applications can easily copy-paste material [51].

In some countries, there is a lot of pressure on researchers to publish so that if they do not publish in journals with high impact factors or other internationally indexed journals, they will not get promoted even if they have high instructional skills. This situation represents the familiar saying "Publish or perish." Therefore, some scholars may make ethical mistakes under the pressure to make progress and to hurry up with publishing [9].

Cultural issues are specially considered in the problem of plagiarism. Cheating and plagiarism is an acceptable practice among the teachers and students of countries where there is little awareness of copyright [52]. A study reveals that students with a stronger belief in detecting plagiarism commit this less than others and turn out to have better writing skills, self-confidence and creativity [53]. Robert Harris takes students' looking for short cuts, their low interest in the research subject, their low planning skills, mismanagement of time, lack of skills in scientific writing and their interest in ignoring regulations as some of the reasons why students take to plagiarism [54].

Another study shows that the following are among the most important reasons why students plagiarize:

1. Genuine lack of understanding. Some students plagiarize unintentionally, when they are not familiar with proper ways of quoting, paraphrasing, citing and referencing and/or when they are unclear about the meaning of 'common knowledge' and the expression 'in their own words'.
2. Efficiency gain. Students plagiarize to get a better grade and to save time. Some cheat because of what Straw (2002) calls 'the GPA thing, so that cheating becomes 'the price of an A' (Whiteman & Gordon, 2001). Auer &Krupar (2001) identify a strong consumer mentality amongst students, who seem to believe that 'they should get grades based on effort rather than on achievement'.
3. Time management. There are many calls on student's time, including peer pressure for an active social life, commitment to college sports and performance activities, family responsibilities and pressure to complete multiple work assignments in short amounts of time. Little wonder that Silverman (2002) concludes that 'students' overtaxed lives leave them so vulnerable to the temptations of cheating'.
4. Personal values/attitudes. Some students see no reason why they should not plagiarize or do it because of social pressure, because it makes them feel good or because they regard short cuts as clever and acceptable.
5. Defiance. To some students plagiarism is a tangible way of showing dissent and expressing a lack of respect for authority. They may also regard the task set as neither important nor challenging.
6. Students' attitudes towards teachers and class. Some students cheat because they have negative student attitudes towards assignments and tasks that teachers think have meaning but they don't (Howard, 2002). Burnett (2002) emphasizes the importance of a relationship of trust between student and teacher, because 'the classes in which [students] are more likely to cheat … are those where students believe their professor doesn't bother to read their papers or closely review their work'.
7. Denial or neutralization. Some students deny to themselves that they are cheating or find ways of legitimizing it by passing the blame on to others
8. Temptation and opportunity. It is both easier and more tempting for students to plagiarize as information becomes more accessible on the Internet and web search tools make it easier and quicker to find and copy.
9. Lack of deterrence. To some students the benefits of plagiarizing outweigh the risks, particularly if they think there is little or no chance of getting caught and there is little or no punishment if they are caught [31].

Some of the perceived obstacles to changing the management of plagiarism are:

a reluctance by staff to process a case of suspected plagiarism due to the time and workload involved in proving‟ the plagiarism;" a reluctance

to become the one who dares to differ where it has been somewhat common practice to "turn a blind eye" to some relatively minor cases of plagiarism; a perception that the University is reluctant to act on suspected plagiarism and that therefore the effort expended by individual staff is likely to be fruitless in terms of dissuading or punishing plagiarism; a fear of risking collegial relationships with students by seeming or becoming authoritarian through a focus on minimizing plagiarism; a concern that following through with cases of repeated plagiarism that may lead to student expulsion might damage the international reputation of the faculty or university; and a further concern that such damage to reputation may result in reduced international enrolments; fear of harassment from the student(s) accused of plagiarism and/or from their friends (such harassment occurred previously in the faculty); fear of student complaints if accusations of plagiarism are made (this had been an issue for some sessional staff who were concerned that a student complaint might mean the end of their employment) [55].

Detecting Plagiarism

Detecting plagiarism is hard and this makes plagiarism a threat to the health of scientific literature. Often plagiarism is recognized by learned reviewers who possess up-to-date knowledge in their own specialist filed [23].

The following include some of the methods that can be used by researchers to detect plagiarism. 1) General sight overview: the academic staff should assess the sentence structure, grammar and idioms used in the students' assignments. They should examine the work which is lower or higher than the student's abilities can afford; 2) Search of online bookstores: these stores help the academic staff to decide whether the students mentioned the right dates for publications or whether the sources used were appropriate to the subject in hand; 3) Search of keywords: searching keywords in search engines is another tool in the hands of academics to find instances if plagiarism, although today's searching technology makes it possible for us to search a whole text, too; 4) a use of plagiarism services: there are many software applications and tools and web sites that can help us detect plagiarized texts [56]. Most of these tools use correlation techniques to detect similarities between documents. Only some of these applications are free and they are all good for English texts. There are methods, however, that can be used in any language. The Glatt plagiarism service, for example, is a computer application which does not depend on correlation techniques. It deletes every fifth word in a paper suspected of plagiarism and the author of the paper is then asked to fill in the missing words. If the student can't fill in 77% of the missing words, the work is most probably plagiarized [17]. Wcopyfind is a free application on the Internet which can be used to detect plagiarism. This software examines a group of document files to compare their content [57].

There are other tools such as http://ithenticate.com, http://www.crossref.org) and http://turnitin.com to discover plagiarism, but these tools can examine the papers indexed in MEDLINE only [42]. This area of study has been attended to by Turnitin and Safe Assign in the last 10 years. Kohler and Weber-wulf carried out a study in 2010 on 47 systems of direct plagiarism detection and concluded that only 5 of them were to some extent useful [58].

There are three approaches to detecting plagiarism. The most common approach is by comparing the document against a number of other documents on a word by word basis. The second approach is by taking a characteristic paragraph and just doing a search with a good search engine like Google. And the third is by style analysis, which is usually called stylometry [30]. Computer applications reports cannot be simply relied here and there will be need for specialist interpretation in such cases [30]. Detecting plagiarism is sometimes very difficult, especially when rephrasing has occurred, when non-electronic sources have been used and when there is shift of language between the original document and the plagiarized one [30]. Although comparing abstracts is a good way to detect plagiarism, a comparison of the full texts will render better results [23].

Systems of retrieving data or plagiarism detecting applications are capable of finding plagiarism where a verbatim copy of words has happened, but what happens when the order of words has been changed but not the overall meaning of the sentences? Naturally, in such cases, the software will not be able to detect plagiarism and the plagiarizer will be able to deceive it. Therefore, these systems

may become useless in the short run [11].

Strategies to Tackle Pagiarism

As Delvin points out universities do not like to endanger their reputations for the sake of plagiarizers [59]. One of the measures needed to assure the quality of universities is to make sure that their assessing policies and activities are useful enough that their assessment is effectively examined in terms of its validity, reliability and fairness. Some of the plagiarism preventing measures recommended by quality assuring organizations include providing a definition of academic misconduct with regard to plagiarism, cheating, identity fraud and using inappropriate content" [60]. In order to effectively and fairly fight against plagiarism, students and staff need to have the same definition of it [61]. A preventing approach on the part of the staff can eradicate this sort of misconduct and make for academic progress and consistency in freshmen. Along with practical approaches based on skills, interactive prevention can not only improve the students' skill in referencing and citing and avoiding plagiarism but also increase their awareness of and sensitivity to this matter [62].

University authorities are responsible for preventing plagiarism in all departments of their universities. The universities' policies in this regard should be clearly defined and announced and disseminated among students and staff, preferably published on the universities' web sites, in libraries, student deputy offices, research centers and dormitories. Academic staff should always talk to students about academic values and avoiding plagiarism. Students, on the other hand, should try to improve their skills in writing papers, research methodology, and organizing data. University teachers should encourage academic honesty in students, clearly define plagiarism for them, and point out to them that they should reference the accessed materials. University policies can also help staff to decide how to deal with plagiarism [17].

Burke points out that universities should focus on teaching student as to how they should avoid plagiarism [63]. The results show that teaching students, especially in the first year, is more effective than other ways of preventing plagiarism. Landau argues that taking an active approach to plagiarism is very important because students who are not fully aware of such misconduct may suppose that they know well about it and not try to learn more. Similarly, the staff may also wrongly think that students are well aware, and thus lose the opportunity to teach students to avoid plagiarism [40]. Although the preventing approach takes more time, it is more effective than other approaches [64]. Exercises and activities encouraged by staff have led to good results. These activities include methods of appropriate citation, quotation, paraphrasing, phrasing and presenting some instances of plagiarism [40].

The attempt by some Australian universities to hide their management of plagiarismhas made it difficult to share the best practices in this field. Although it seems that common policies are in practice with respect to plagiarism, there is nothing to indicate the success of these policies. Delvin reports that for some, "catch and persecute" leads to a decrease in plagiarism [59] while there is little evidence to suggest the effectiveness of such a measure [52]. Gallant similarly argues that traditional methods of preventing plagiarism, such as persecution as a preventing measure, honor code systems and instructor detection are not effective today [8]. In the UK, in order to minimize plagiarism, they makes use of special courses, assessment, giving awareness to students, teaching student the necessary skills, detecting plagiarism, persecution and special policies, and instructional programs [65].

Conway and Groshek have shown that student ethics is flexible and can be shaped at any level of education. Students showed in this research that they are concerned about ethical violations and expect that severe punishment will be considered for those students who plagiarize or fabricate materials." Repeating anti-plagiarism actions at any stage of education can empower students [66]. MIT has defined good methods and policies to manage academic misconduct. In this university, teachers are rewarded for teaching the right academic behavior to students [67]. The Online Writing and Communication Centerof this university runs a program for improving students' writing skills and explains different aspects of plagiarism [68]. In some universities, such as Berkeley, people are academically sanctioned for plagiarism [69]. In Stanford University, students learn about the university policies to deal with academic misconduct, copyright and

fair use of materials.

Although there is always need for good inspections, the responsibility to keep research integrity lies with the scientific community itself and academic staff should make sure that students learn about this integrity. Authors should guarantee that their reported work is new and correct. Scholars who agree to review papers should feel responsible for doing informed, thorough and conscientious reviewing. Journal editor, who are themselves distinguished scholars, should assure the originality of the material they publish [42]. The ideas and thoughts of different thinkers and authors are inevitable connected. So, it is a great responsibility of authors to make sure that no plagiarism occurs when they publish their results. This means that the authors must do whatever they can to ensure that the words of their papers are theirs. They should be always sure that it is clear for their readers whether the ideas presented in the papers are theirs or others' and this could be clear by citing earlier published sources [24].

The process of peer-reviewing is the best mechanism to ensure the high quality of publications. But recent studies have shown that lack of appropriate standards can result in duplicate publication as well as publication of papers which include plagiarism [42]. At present, plagiarism tackling approaches focus on instructions to students and making them aware of the related policies and possible outcomes. For instance, students are taught to utilize to access and use sources in the right way. Also, developing scientific integrity and honor code systems are among good approaches to plagiarism [70]. Carroll argues that teachers should focus on prevention [71]. McCabe similarly thinks that reducing the chances of plagiarism is an important tool in reducing scientific misconduct [72].

Authors should bear in mind that it is not acceptable to republish a paper which has already been published, but this rule has the following exceptions, if the right disclosure is made to the editors and reader:

Prior publication in abstract form only (generally <400 words);

A study is too large and/or complex to be reported in one article. A proposed rule of thumb is an expansion of the original article by 50%. However, each article should address a different distinct and important question;

Competing submissions of coworkers who disagree on analysis and interpretation of the same study;

Articles from different groups of authors who have analyzed the same data. This is often the case with very large administrative data sets or large national surveys sponsored by government agencies;

Republication of an article in another language with cross-referencing. There are mixed thoughts on the acceptability of this practice. Typically the two (or more) journals need to work together and often permission to publish is needed. The International Council of Medical Journal Editors has published criteria for this practice. While publication of data in an uncommon language need not necessarily prevent it being presented in English, secondary publication should follow the International Council of Medical Journal Editors guideline in the uniform requirements [9].

Strategies to Avoid Plagiarism

1. Read the instructions for authors provided by the journal.
2. Always acknowledge the contributions of others and the source of ideas and words, regardless of whether paraphrased or summarized.
3. Use of verbatim text/material must be enclosed in quotation marks.
4. Acknowledge sources used in the writing.
5. When paraphrasing, understand the material completely and use your own words.
6. When in doubt about whether or not the concept or fact is common knowledge, reference it.
7. Make sure to reference and cite references accurately.
8. If the results of a single complex study are best presented as a cohesive whole, they should not be sliced into multiple separate articles.
9. When submitting a manuscript for publication containing research questions/hypotheses, methods, data, discussion points, or conclusions that have already been published or disseminated in a significantmanner (such as previously published as anarticle in a separate journal or a report posted on theInternet), alert the editors and readers. Editorsshould be informed in the cover letter, and readersshould be alerted by highlighting and citing theearlier published work.
10. When submitting a manuscript for potential

publication, if there are any doubts or uncertainty about duplication or redundancy of manuscripts originating from the same study, the authors should alert the editors of the nature of the overlap andenclose the other manuscripts (published, in press/ submitted, unpublished) that might be part of them an uscript under consideration. Augmenting old data that was previously published with new additional data and presenting it as a new study can be an ethical breach and should be fully disclosed to the editors.

11. Write effective cover letters to the editor, especially regarding the potential for overlap in publication. The cover letter should detail the nature of the overlap and previous dissemination and ask for advice on the handling of the matter.

12. Become familiar with the basic elements of copyright law [28].

Conclusion

Ethical problems in science are quickly increasing and have become controversial issues in universities and educational research institutes. These problems have also been reflected in media news recently. The growth of information technology, competition between countries, rapid growth of knowledge, fast multiplication of scientific journals, lack of good explication of plagiarism and different understandings of it, lack of awareness, mismanagement of time, and low culture etc. have all contributed to the prevalence of plagiarism in the scientific community. This has worried scientific institutes and has made them react to it. Some institutes focus on detecting and persecuting while others concentrate on preventions and teaching the right behavior. Excessive stress on detection of plagiarism has made for the development of data retrieving systems in recent years, but these are not effective enough, and even if they were, they would not be the best solutions. Effective prevention through proper education at the right time, proper interaction between teachers and students and devising appropriate policies for this purpose are possible means of tackling plagiarism.

References

1.Tavakol M, rad MN. Plagiarism with explanation of the sociology of science .Journal of Ethics in Science and Technology. 2010;4(3):1-16.

2.DeVoss D, Rosati AC. “It wasn’t me, was it?” Plagiarism and the Web. Computers and Composition. 2002;19(2):191-203.

3. Pakjou A, Izadi M, Masoudipour Sh, Fazel M. “Pattern of travel medicine ethics in international cooperation programs.” Journal of Military Medicine. 2011;13(2):117-123.

4.Stewart Jr CN. Research Ethics for Scientists: A Companion for Students: Wiley; 2011.

5.Roberts DM, Toombs R. A scale to assess perceptions of cheating in examination-related situations.Educational and Psychological Measurement. 1993;53(3):755-62.

6.Brown GO. Out of the way: how the next copyright revolution can help the next scientific revolution. PLoS Biol. 2003 Oct;1(1):E9.

7.Beheshti SM. Plagiarism. Tehran 2011 [Feb, 7]; Available from: http://www.tebyan.ne.

8.Gallant TB. Academic integrity in the twenty-first century: A teaching and learning imperative: Jossey-Bass Inc Pub; 2008.

9.Miziara ID. Ethics in scientific publications: the double copyright problem. Braz J Otorhinolaryngol. 2010 Sep-Oct;76(5):543.

10.Hart M, Friesner T. Plagiarism and poor academic practice–a threat to the extension of e-learning in higher education? Electronic Journal on e-learning. 2004;2(1):89-96.

11.Foudeh P. Performance evaluation of information retrieval methods in dealing with plagiarism. Scientific Communication. 2009;2(3):1-5.

12.Mallon T. Stolen words: Forays into the origins and ravages of plagiarism: Penguin Books; 1991.

13.loghatnamehDehkhoda. Tehran2012 [cited 2012 May, 12]; Available from: http://www.jasjoo.com/books/wordbook/dehkhoda.

14.Bartlett J. Familiar quotations: a collection of passages, phrases, and proverbs traced to their sources in ancient and modern literature: Little, Brown, and company; 1904.

15.Barnhart RK, Steinmetz S. Chambers Dictionary of Etymology–the origins and development of over 25,000 English words.Chambers. Edinburgh, uk; 1988.

16.Fialkoff F. There’s no excuse for plagiarism. Library Journal. 1993;118((17)):56.

17.Austin MJ, Brown LD. Internet plagiarism: Developing strategies to curb student academ48.Ben

ic dishonesty. Internet and Higher Education. 1999;2:21-34.
18.Conners M. Cybercheating: the Internet could become the newest battleground in academic fraud. The Muse. 1996.
19.College LHU. Undergraduate modular scheme (awards of the University of Liverpool)—appendices. 2012 [cited 2012 May, 15]; Available from: www.hope.ac.uk/compass/.
20.Pyer H. Plagiarism. 2012 [cited 2012 May,16]; Available from: at: http://online.northumbria.ac.uk.
21.Stebelman S. Cybercheating: Dishonesty goes digital. American Libraries. 1998;29(8):48-50.
22.Vessal K, Habibzadeh F. Rules of the game of scientific writing: fair play and plagiarism. The Lancet. 2007;369(9562):641.
23.Sharma BB, Singh V. Ethics in writing: Learning to stay away from plagiarism and scientific misconduct. Lung India. 2011 Apr;28(2):148-50.
24.Mason PR. Plagiarism in scientific publications. J Infect DevCtries. 2009;3(1):1-4.
25.Dellavalle RP, Banks MA, Ellis JI. Frequently asked questions regarding self-plagiarism: How to avoid recycling fraud. J Am Acad Dermatol. 2007 Sep;57(3):527.
26.Hayes N, Introna LD. Cultural values, plagiarism, and fairness: When plagiarism gets in the way of learning. Ethics & Behavior. 2005;15(3):213-31.
27.Kravitz RL, Feldman MD. From the Editors Desk: Self-Plagiarism and Other Editorial Crimes and Misdemeanors. J Gen Intern Med. 2010 Nov 9.
28.Cicutto L. Plagiarism: avoiding the peril in scientific writing. Chest. 2008 Feb;133(2):579-81.
29.Habibzadeh F, Shashok K. Plagiarism in scientific writing: words or ideas? Croat Med J. 2011 Aug 15;52(4):576-7.
30.Maurer H, Kappe F, Zaka B. Plagiarism-a survey. Journal of Universal Computer Science. 2006;12(8):1050-84.
31.Park C. In other (people›s) words: Plagiarism by university students literature and lessons.Assessment & Evaluation in Higher Education. 2003;28(5):471-88.
32.Roig M. Plagiarism and paraphrasing criteria of college and university professors. Ethics & Behavior. 2001;11(3):307-23.
33.Park C. Rebels without a clause: towards an institutional framework for dealing with plagiarism by students. Journal of further and higher education. 2004;28(3):291-306.
34.University BGS, Administrators NAoSP, Affairs AfSJ, Integrity CfA. Academic Integrity: The Truth of the Matter: Bowling Green State University; 1995.
35.Carroll J. A handbook for deterring plagiarism in higher education: Oxford Centre for Staff and Learning Development Oxford; 2007.
36.Kellogg AP. Students plagiarize online less than many think, a new study finds. Chronicle of Higher Education. 2002;48(23):44.
37.Satterthwaite T. Catching the cut-and-pasters. 2003 [cited 2003 May,18]; Available from: http:///www.here.ac.uk/inside_he/archive/catching_the_cut_and_past4311.cfm?archive=yes.
38.CMJ S. Prevalence of plagiarism among medical students. Croat Med J. 2005;46(1):126-31.
39.Dordoy A, editor. Cheating and plagiarism: student and staff perceptions at Northumbria 2002.
40.Landau JD, Druen PB, Arcuri JA. Methods for helping students avoid plagiarism. Teaching of Psychology. 2002;29(2):112-5.
41.Carroll J. Suggestions for teaching international students more effectively. Learning and Teaching Briefing Papers Series, Oxford Brookes University Retrieved April. 2002;1:2005.
42.Long TC, Errami M, George AC, Sun Z, Garner HR. Responding to possible plagiarism. Science. 2009;323(5919):1293-4.
43.Rode Sde M. Plagiarism in scientific publication. Braz Oral Res. 2011 Apr;25(2):101.
44.Ashworth P, Bannister P, Thorne P, Unit SotQRMC. Guilty in whose eyes? University students perceptions of cheating and plagiarism in academic work and assessment.Studies in Higher Education. 1997;22(2):187-203.
45.Flint A, Clegg S, Macdonald R. Exploring staff perceptions of student plagiarism. Journal of further and higher education. 2006;30(02):145-56.
46.Auer NJ, Krupar EM. Mouse click plagiarism: The role of technology in plagiarism and the librarian›s role in combating it. Library Trends. 2001;49(3):415-32.
47.Macdonald R. Why don't we turn the tide of plagiarism to the learners' advantage.Times Higher Educational Supplement. 2000;24.

48.Benning V. Higher learning, lower behavior-students cheat by computer. The Seattle Times. 1998:A7.
49.Angélil-Carter S. Stolen language? Recherche. 2000;67:02.
50.Hyland F. Dealing with plagiarism when giving feedback. ELT Journal. 2001;55(4):375-81.
51.Grover D. Plagiarism and the internet: the use of correlation techniques to detect plagiarism. Computer Law and Security Report. 2003;19(10):6-8.
52.Hamilton M. Managing student plagiarism in large academic departments. 2002.
53.Martin DF. Plagiarism and technology: A tool for coping with plagiarism. The Journal of Education for Business. 2005;80(3):149-52.
54.Harris R. Anti-plagiarism strategies for research papers. Virtual salt. 2004;4:6.
55.Devlin M, editor. Policy, preparation, prevention and punishment: one faculty's holistic approach to minimising plagiarism2011: Swinburne University.
56.Ryan J. Student plagiarism in an online world. ASEE Prism Magazine. 1998.
57.Bloomfield L. WCopyfind 2.1 instructions. 2012 [cited 2012 May, 13]; Available from: http://www.plagiarism.phys.virginia.edu/WCopy.find%202.1.html.
58.Sheard J, Dick M. Directions and Dimensions in Managing Cheating and Plagiarism of IT Students. 2012.
59.Devlin M. Problem with plagiarism. Campus review. 2003;12(44):4-5.
60.QAA. Code of practice for the assurance of academic quality and standards in higher education. 2000 [cited 2012 May, 15]; Available from: www.qaa.ac.uk/public/Cop/COPaosfinal/contents.htm.
61.Stefani LAJ, Carroll J, Centre LG. A briefing on plagiarism: Learning and Teaching Support Network; 2001.
62.Dunn K. Recommendations for an Interactive Approach to Plagiarism Prevention. 2011.
63.Burke M. Deterring plagiarism: A new role for librarians. Library Philosophy and Practice (e-journal).2005:10.
64.Compton J, Pfau M. Inoculating against pro-plagiarism justifications: Rational and affective strategies. Journal of Applied Communication Research. 2008;36(1):98-119.
65.Mitchell CM, Wisbey ME. Cheaters Never Prosper, but Do They Get College Degrees? College Student Affairs Journal. 1995;15(1):87-93.
66.Conway M, Groshek J. Ethics Gaps and Ethics Gains: Differences and Similarities in Mass Communication Students' Perceptions of Plagiarism and Fabrication. Journalism & Mass Communication Educator. 2008;63(2):127-45.
67.Massachusetts Institute of Technology Policies and Procedures. 2012 [cited 2012 May, 23]; Available from: http://web.mit.edu/policies/10.0.html.
68.MIT Online Writing and Communication Center. 2012 [cited 2012 May, 23]; Available from: http://web.mit.edu/writing.
69.Berkeley University of California, Student Conduct, Sanctions. 2012 [cited 2012 May, 23]; Available from: http://students.berkeley.edu/osl/sja.asp?id=1004.
70.Barrett R, Malcolm J. Embedding plagiarism education in the assessment process. International Journal for Educational Integrity. 2006;2(1):38-45.
71.Carroll J. Institutional issues in deterring, detecting and dealing with student plagiarism. Joint Information Systems Committee (JISC). 2004.
72.McCabe DL, Treviño LK, Butterfield KD. Cheating in academic institutions: A decade of research.Ethics &Behavior. 2001;11(3):219-32.

Plagiarism: Taxonomy, Tools and Detection Techniques

Hussain A Chowdhury and Dhruba K Bhattacharyya

Dept. of CSE, Tezpur University

Abstract

To detect plagiarism of any form, it is essential to have broad knowledge of its possible forms and classes, and existence of various tools and systems for its detection. Based on impact or severity of damages, plagiarism may occur in an article or in any production in a number of ways. This survey presents a taxonomy of various plagiarism forms and include discussion on each of these forms. Over the years, a good number tools and techniques have been introduced to detect plagiarism. This paper highlights few promising methods for plagiarism detection based on machine learning techniques. We analyse the pros and cons of these methods and finally we highlight a list of issues and research challenges related to this evolving research problem.

Keywords: Intrinsic, Extrinsic, Elsevier, Nearest-neighbour, Textual plagiarism, Source-code plagiarism

1. Introduction

Due to the digital era, the volume of digital resources has been increasing in the World Wide Web tremendously. Today, creation of such digital resources and their storage and dissemination are simple and straight forward. With the rapid growth of these digital resources, the possibility of copyright violation and plagiarism has also been increasing simultaneously. To address this issue, researchers started working on plagiarism detection in different languages since 1990. It was pioneered by a copy detection method in digital documents [1]. However, the software misuse detection was initiated even much earlier, in 1970 by detecting plagiarism among programs [2]. Since then, a good number of methods and tools have been developed on plagiarism detection which are available online. But it is very much chaotic when one wants to choose the best plagiarism detection method or plagiarism detection tool. It may be due to lack of controlled evaluation environment in plagiarism detection research. Plagiarism is the presentation of another's words, work or idea as one's own [3]. It has two components, viz., (1) Taking the words, work or ideas from some source(s) and (2) Presenting it without acknowledgments of the source(s) from where words, works or ideas are taken [3]. Plagiarism can appear in different forms in an articles. However, there are mainly two types of plagiarisms typically found to occur, such as (1) textual plagiarism and (2) source code plagiarism [4]. Plagiarism may occur within same natural language or it may appear between two or more different languages. Many researchers or software companies still trying to provide an efficient method or tool for plagiarism detection. There are mainly two types of plagiarism detection approaches available based on whether external resources or references are used or not during plagiarism detection, such as (1)

intrinsic plagiarism detection, where no external references are used and (2) extrinsic plagiarism detection, where external references are used [5].

In the yester years, a good number of tools and techniques have been introduced to detect plagiarism of various forms. Several efforts have been made [6] [5][7][8] to survey these works. However, unlike other surveys, this survey is attractive in view of the following points.

- 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.
- Finally, in includes a list of issues and research challenges.

The rest of the paper is organized as follows: Section 2 presents fundamentals of plagiarism and its classification. It reports a taxonomy of various known plagiarism forms. In section 3 we discuss fundamentals of detection and a large number of detection techniques. Section 4 reports a list of tools for plagiarism detection and discuss their features. Section 5 presents a list of issues and research challenges. Finally, section 6 draws the conclusions of this paper.

2. Plagiarism and Its Types

As stated in [3], 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.

Plagiarism can appear in different forms in a document, work, production or program. Two basic types of plagiarisms [9] are (a) Textual plagiarism and (b) Source Code plagiarism.

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

Textual plagiarism is commonly seen in education and research. Figure 1 (a) shows an example of textual plagiarism where entire word-for-word are taken from source without direct quotation. Textual plagiarism further can be divided into seven sub categories based on its forms and application [4][10] as shown in Figure 2. 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 on two ways as given below.
    - 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.
    - 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.
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.

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. In general, intrinsic plagiarism detection methods can detect paraphrasing, idea, and mosaic textual plagiarism sub-types [4], whereas, external plagiarism detection methods can detect clone, metaphor, retweet and possibly (with a low probability or none) self plagiarisms and error 404 [4]. However, in the recent developments, such demonstration is not very prominent between these two types of detection methods.

In *source code plagiarism*, 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 plagiarism where entire program is represented again in different way by changing syntax. Typically it is seen, in educational institutes. This type of plagiarism, as shown in Figure 2 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.

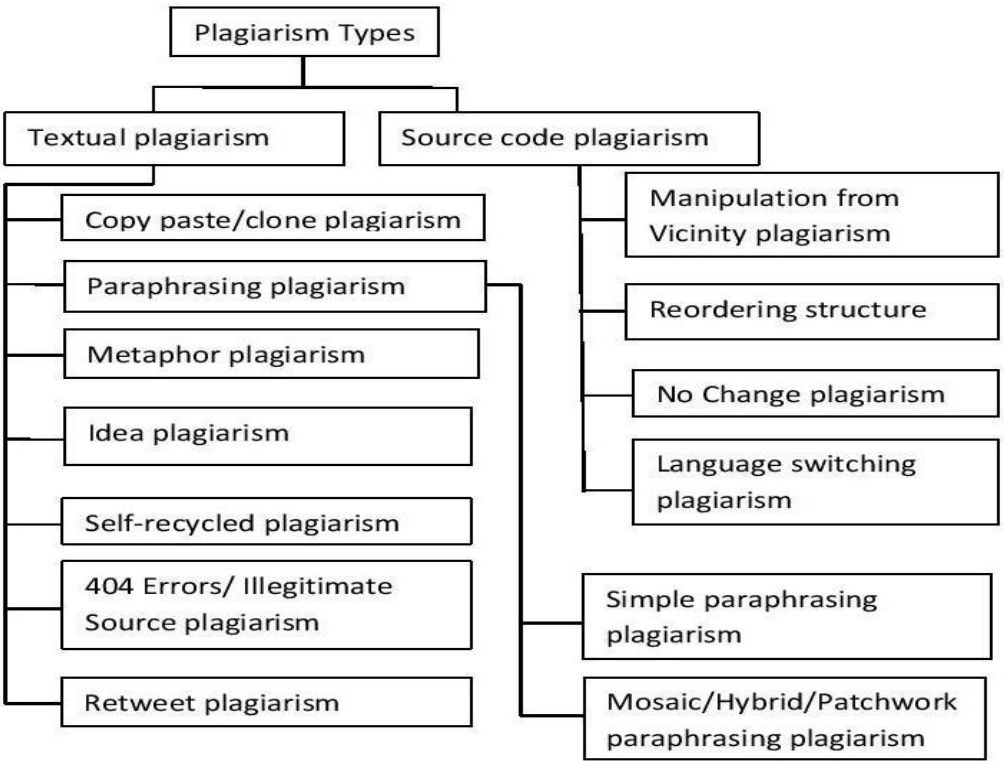

Figure 2: Taxonomy of plagiarism

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.
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 declares it as his/her own.

3. Plagiarism Detection

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 [5] 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.
    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. }

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.

In Figure 3 a schematic view of the basic plagiarism detection has been shown for both text documents and source codes. It accepts an input candidate text document and attempts to identify the text segments in the document plagiarized from some sources. It can be in monolingual as well as in cross-lingual framework. Textual plagiarism detection can be classified based on how textual features are used to characterize documents. 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. These textual features are used in both extrinsic and intrinsic as well as in cross-lingual plagiarism detection.

In case of source code plagiarism detection, human intervention is required to detect plagiarism. 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 [11].

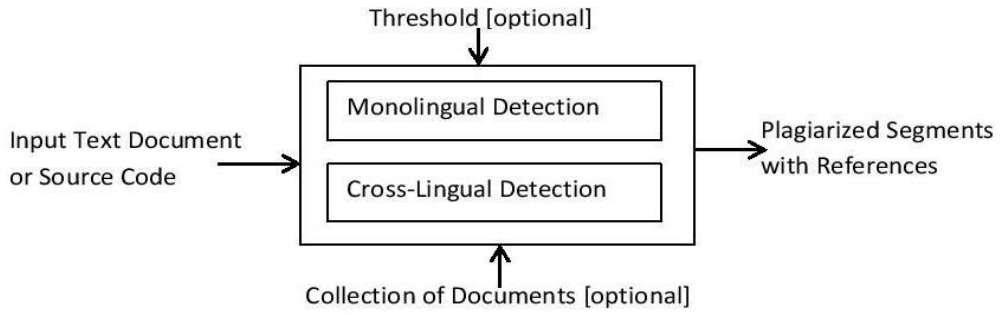

Figure 3: Basic plagiarism detection system

3.1 Similarity Measures for Comparing Documents or text segments

To detect plagiarism we have to measure similarity between two documents. We observe that most researchers use the following two types of similarity metrics.

1. *String Similarity Metric:* This method is commonly used by extrinsic plagiarism detection algorithms. Hamming distance is a well-known example of this metric which estimates number of characters different between two strings x and y of equal length, Levenshtein Distance [12][13] is another example, that defines minimum edit distance which transform x into y, similarly, Longest Common Subsequence [14][2] measures the length of the longest pairing of characters between a pair of strings, x and y with respect to the order of the characters.
2. *Vector Similarity Metric:* Over the decade, a good number of vector similarity metrics have been introduced. A vector based similarity metric is useful in calculating similarity between two different documents. Matching Coefficient [15] is such a metric that calculates similarity between two equal length vectors. Jaccard Coefficient [16] is author such metric used to define number of shared terms against total number of terms between two identical vectors, Dice Coefficient [17] is similar to Jaccard but it reduces the number of shared terms, Overlap Coefficient [18] can compute

similarity in terms of subset matching, Cosine Coefficient [19] to find the cosine angle between two vectors, Euclidean Distance the geometric distance between two vectors, Squared Euclidean Distance places greater weight on that are further apart, and Manhattan Distance can estimate the average difference across dimensions and yields results similar to the simple euclidean distance.

3.2 Plagiarism Detection Methods

Detection of plagiarism in text document with high accuracy is a challenging task. In the past two decades, a large number of methods have been reported by researchers to handle this task. These methods can be classified into eleven distinct categories. Some prominent methods under each of these categories are discussed next. Also, we have analysed their pros and cons, and reported in a tabular form in Table 1.

1. *Character-Based Methods:* Most plagiarism detection methods belong to this category. These methods exploit character-based, word-based, and syntax-based features. It utilizes these features to find similarity between a query document and existing documents. However, the similarity between a pair of documents may be estimated using both exact matching and approximate matching. In exact matching, every letter in both the strings must be matched in the same order. Our survey reveals that most detection techniques are developed based on n-gram or word n-gram based exact string similarity finding approach. For instance, Grozea et al. [20] use character 16-gram matching, whereas the authors of [21] use word 8-gram matching. Similarly, some researcher has made an effective use of approximate string matching approach. This string matching shows degree of similarity/dissimilarity between two strings. There are several proximity measures available to support the approximate string matching. One can use string similarity metric or vector similarity metric for the purpose.
2. *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. Our observation is Cosine coefficient and Jaccard coefficients are popular and effective in finding similarity between two vectors. Cosine coefficient in detecting partial plagiarism without sharing documents content. Hence it is useful to detect plagiarism in documents where submission is considered as confidential [22].
3. *Syntax-Based Methods*: These methods exploit syntactical features like part of speech (POS) of phrase and words in different statements to detect plagiarism. The elements of basic POS tag are verbs, nouns, pronouns, adjectives, adverbs, prepositions, conjunctions and interjections. In [14] [13], the authors use POS tag features followed by string similarity metric to analyse and calculate similarity between texts. The authors of [23] use syntactical POS tag to represent a text structure as a basis for further comparison and analysis i.e., documents containing same POS tag features are carried out for further analysis and for identification of source of a plagiarism.
4. *Semantic-Based Methods:* A sentence may be defined as an ordered group of words. Two sentences may be same but the order of their words may be different. In Figure 1 (a), sentence is constructed by just transforming from active voice to passive voice but the semantics of the sentences are same. WordNet [24] is used in this content to find the semantic similarity between words or sentences. The degree of similarity between two words used in knowledge-based measures by Gelbukh [25] is calculated using information from a dictionary. This similarity between two words is used as semantic similarity between two words. In another approach, Resnik [26] used WordNet to calculate the semantic similarity, whereas, Leacock's et al., [27] determine semantic similarity by counting the number of nodes of shortest path between two concepts.

5. *Fuzzy-Based Methods:* In a fuzzy-based method, similarity of text such as sentences is represented by values ranging from zero (entirely different) to one (exactly matched). Here, the words in a documents 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 [28]. This method is attractive because it can detect similarity between documents with uncertainty. In [28], a correlation matrix is constructed which consists of words and their corresponding correlation factors that measures the degree of similarity among different words. Then, it obtains the degree of similarity among sentences by computing the correlation factors between pair of words from two different sentences in their respective documents. In [29], the degree of similarity of two documents or any two Web documents are identified by using fuzzy IR approach. The authors introduce a tool for this purpose. There is another method discussed in [30] which adapts fuzzy approach to find in what extent two Arabic statements are similar. For that they used a plagiarism corpus of 4477 sources statements and 303 query/suspicious statements.
6. *Structure-Based Methods:* Unlike those methods above, developed based on lexical, syntactic, and semantic features of the text in documents to find similarity between two documents, a structure based method uses contextual similarity such as how the words are used in entire documents. However, our survey can find a few methods of this category. Contextual information is generally handled using tree-structure feature representation as can be found in ML-SOM [31]. In [32], the author detects plagiarism in two steps. First step performs document clustering and candidate retrieval using tree-structure feature representation and second step detects by utilizing ML-SOM.
7. *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 [33]. A writer specific style is mostly with author's vocabulary strength or complexity of presenting a document. On the other hand, a reader specific style deals with how a reader can easily understand the texts. One can find usefulness of outlier mining to detect plagiarism in a document under this approach. A detail discussion on Stylometric-Based methods is available in [34].
8. *Methods for Cross-Lingual Plagiarism Detection*: Cross-lingual plagiarism detection is a challenging task. It requires in depth knowledge of multiple languages. Finding appropriate similarity metric for such method is also an important issue. 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 method, and (3) cross-lingual dictionary based methods [5]. A detail survey on Cross-Lingual methods is done in [35]. In [20], 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 [36].
9. *Grammar Semantics Hybrid Plagiarism Detection Methods:* These methods are effective method in plagiarism detection for their use of natural language processing. They are capable of detecting copy/paste and paraphrasing plagiarism accurately. Such methods eliminate the limitations of semantic-based method. 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 [37][5].
10. *Classification and Cluster-Based Methods*: In information retrieval process, supervised and unsupervised grouping of documents plays an important role. In many research problem such as text summarization [38], text classification [39],

Table 1: PLAGIARISM DETECTION TECHNIQUES: A General Comparison

Author & Name	Intrinsic(I)/Extrinsic	Approach used	mode 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

and plagiarism detection [40], classification and clustering are useful in reducing the search space during the information retrieval process. It helps in reducing the document comparison time significantly during plagiarism detection. Some methods [41][42] use keywords or specific words to cluster the similar sections of documents.

11. *Citation-Based Methods:* In [43], a novel method is proposed to detect plagiarism in citation basis. This method is a new approach towards detecting plagiarism and scientific documents that have been read but not cited. Citation-based methods belong to semantic plagiarism detection techniques because these techniques use semantics contained in the citation in a document [44]. The similarity between two documents is computed based on the similar patterns in the citation sequences [44].

4. Plagiarism Detection Tools

In the past two decade, several plagiarism detection tools have been developed. Some of these tools are discussed in brief, next. Also, we have analyzed their pros and cons, and reported in a tabular form in Table 2 We reported the classification of tools in Figure 4

i. **SafeAssignment [6*]*:** This anti-plagiarism checker claims to search an index of 8 billion documents available in the Web. It uses some major scholastic databases like ProQuest™, FindArticles™ and Paper Mills during searching and detection process. SafeAssignment maintains a database where user account is essential to keep fingerprints of the submitted documents in order to avoid any legal or copy right problem. This tool uses proprietary searching and ranking algorithms for match detection of fingerprints with its resources. The results of plagiarism detection is presented to the user within couple of minutes.

ii. **Docol©c[6]:** This Web based service uses capabilities like searching and ranking of Google API. The submitted document is uploaded to a server and evaluation is done in the server side. The software provides a simple console to set fingerprint (search fragments) size, date constraints, filtering and other report related options. The evaluation result is sent to the user through email identifying plagiarized sections and sources of plagiarism. This is totally Google API dependent and so it may be unavailable at any point of time.

iii. **Urkund [6]:** This is another Web based service which carry out plagiarism detection in server side. This is an integrated and automated solution for plagiarism detection. This is a paid service which uses standard email system for document submission and for viewing results. This system claims to process 300 different types of document submissions and it searches through all available online sources. It gives more priority to educational sources of documents more during searching.

iv. **Copycatch [6]:** This is a client-based tool which utilizes the local database of documents during comparison. It offers ‘gold’ and ‘campus versions’, providing comparison capabilities against large repository of local resources. It has another Web version which utilizes the capabilities of Google API for plagiarism detection across the Internet. To use the Web version, user needs personal Google API licence through signup.

v. **WCopyfind[6]:** It is an open source plagiarism detection tool for detection of words or phrases of defined length within a local repository of documents. Its extended version has the capabilities of searching across the Internet using Google API to check plagiarism online.

vi. **Eve2 (Essay Verification Engine [6][45] :** This system is installed in user's computer and it checks plagiarism of a document against Internet sources. It does not contact any online database. It accepts text in several formats but internally converts the input file into text for processing. It presents the user with a report identifying matches found in the Web.

vii. **GPSP - Glatt Plagiarism Screening Program [6]:** This system uses different approaches unlike other mentioned services. It finds and uses the writing style of the author(s) to detect plagiarism. This service works locally and it asks the author to go through a test by filling the blank spaces. The number of correctly filled spaces and time taken to complete the test are used to make a hypothesis about plagiarism. This system is basically developed for teachers and it cannot detect source code plagiarism.

viii. **MOSS - a Measure of Software Similarity [46]:** This system is used to detect source code plagiarism. This service takes batches of documents as input and attempts to present a set of HTML pages to specify the sections of a pair of documents where matches detected. The tool specializes in detecting plagiarism in C, C++, Java, Pascal, Ada, ML, Lisp, or Scheme programs.

ix. **JPlag [47][6]:** It is a Web based source code plagiarism detection tool started in 1997. The tool accepts a set of programs as input to be compared and to present a report identifying matches. JPlag carry out programming language syntax and structure aware analysis to find results. It can detect plagiarism in Java, C and C++ programs. The execution time of this service is less than one minute for submissions of 100 programs of several hundred lines each.

x. **Copyscape [48]:** This system takes URL as input and search for copies of a Web page in the Internet. Copyscape helps to find sites that have copied from someone's Web page content without permission. It has both free and premium version and it pushes the free users to buy their premium by limiting the search features.

xi. **DOC Cop [49]:** This plagiarism detection system creates report displaying the correlation and matches between documents or between documents and the Web. It is free plagiarism detection system. \

xii. **Ephorus [46]:** To access this tool, user is to register with the Ephorus site. Hence, no downloads or installation is needed. The search engine compares a text document to millions of others on the Web and reports back with an originality report [50]. This tool can be freely tried but license needs to be purchased. It is well known in many European universities and organization.

xiii. **ithenticate [51][46]:** This is a successful Web based plagiarism detection tool for any text document. This tool is not required to install in client computer. This application compares input documents against the document sources available on the Web. This well-known tool is used by most well-known journal publishers. It is a easy to use, quick plagiarism checker for professionals. It is designed to be used by institutions rather than personal, but lastly they provided a limit service for single plagiarism detection user like master and doctoral students and this allows them to check a single document of up to 25,000 words.

xiv. **Plagiarism Detect [46]:** To use this tool, user needs to register by providing correct information. After registration, users are allowed to input text in a given text box or as a file by uploading for analysis. This is a free service which finally sends evaluation report to the user's email account with a list of links from where information are copied. It also specifies amount of plagiarism (in \%) detected. User needs to download and install the software in order to use it.

xv. **Exactus Like [52]:** This plagiarism detection system is not able to find simple copy-paste plagiarism but also can detect moderately disguised borrowing (word/phrase reordering, substitution of some words with synonyms [52]. To do this, the system leverages deep parsing techniques. This Web based tool supports most of the popular file formats such as Adobe PDF, Microsoft Word, RTF, ODT and HTML. Currently Exactus Like includes about 8.5 million indexed documents. Internally this tool is basically a distributed system and a demo version of this tool is available online.

xvi. **DupliChecker [53]:** It is a free online plagiarism checker. This tool can be accessed by unregistered user only once, but registered user can check for plagiarism for 50 times in a day. The input file must contain more than 1000 words per similarity

search. User can check content's originality by number of ways such as via copy paste, uploading file or by submitting URL.

xvii. **Plagiarisma [54]:** It is free and simple plagiarism checking tool. This software supports 190+ languages and it does not store any scanned content. The input file can be provided in three ways (1) Copy paste (2) Check by entering URL and (3) Uploading file. However, the tool lacks of advanced features so it cannot be relied for heavy scanning works.

xviii. **Plagiarism Checker [48]:** It was first available in early 2006. This freely available online service uses Google or Yahoo service to check whether documents submitted by students are copied from Internet material or not. It simply encloses each phrase in quotation marks and inserts an OR between each phrase during checking.

xix. **Plagium [55][8]:** This simple plagiarism detection tool, is effective in comparison to many of its counterpart, both in terms of results and algorithm. Though Plagium can be used free to some extent using quick search, their paid version has added benefits such as timeline feature and alert feature which pops up whenever someone's content is plagiarised. This tool has flexibility in pricing option like we can buy search credit either as prepaid plans or monthly plans. This tool allows user to check for plagiarism upto 5000 words without signing up.

xx. **PlagTracker [56]:** It is a popular plagiarism checker for students, teachers, publishers and Website owners. It has a large database of academic publications in million and provides detail report of the scanned work. If someone wants to check assignments in bulk, it requires to subscribe monthly. This tool found useful to ensure whether a test document is plagiarized or not.

xxi. **Quetext [57]:** It uses Natural Language Processing and Machine Learning to detect plagiarism. It performs first internal plagiarism checking and then it goes for external checking. This free tool uses every possible factor for each word to detect plagiarism. It provides support to multiple languages and one can search for unlimited words. To check plagiarism with this tool, one needs just plain copy paste of the text document. The main disadvantage of this tool is that it does not provide detail report. Also it is not user friendly.

xxii. **Turnitin [46][58]:** This an another successful Web based tool provided by iParadigms. The user is needed to upload test document to the system database for plagiarism check the system creates a fingerprint of the document and stores it. In this tool, detection and report generation is carried out remotely. Turnitin is already accepted by 15,000 Institutions and 30 Million Students due to easy to use interface, support of large repository, detailed text plagiarism check and well organized report generation. It can be considered as one of the best plagiarism checkers for teachers.

xxiii. **Viper [8]:** This free plagiarism scanner scans the submitted documents against 10 billion sources and documents present in a computer. It gives peace of mind regarding any accidental plagiarism. This tool offers unlimited resubmitting of documents and it provides links to plagiarised work in the reports.

xxiv. **Maulik [59]:** Maulik detects plagiarism in Hindi documents. It divides the text into n-grams and then matches with the text present in the repository as well as with documents present online. It uses Cosine similarity for finding the similarity score. Maulik is capable of finding plagiarism if root of a word is used or a word is replaced by its synonyms. This tool is superior than existing. Hindi plagiarism detection tools such as Plagiarism checker, Plagiarism finder, Plagiarisma, Dupli checker, and Quetext.

xxv. **Plagiarism Scanner [8]:** This is a fast and effective plagiarism detection tool for students, instructors, publishers, bloggers since 2008. It is a user-friendly online tool. This tool conducts through an in-detail detection for plagiarism of a submitted document within a few minute only. This tool runs against all Internet resources, including Websites, digital databases, and online libraries (such as Questia, ProQuest, etc). It generates a full report, indicating the overall originality rating and the percentage of plagiarized materials in the submitted text. It also provides

Table 2: PLAGIARISM DETECTION TOOLS: A General Comparison

Tool Name & Author	Year	Extrinsic(E)/Intrinsic(I)	User friendly	Submission of single/multiple Files?	Source code availability?	Source(Ref)
SafeAssignment by Mydropbox	2008	E	yes	single	No(Free)	http://www.safeassign.com/
Docolc by IFALT9	2005	E	yes	single	No(Free)	https://www.docoloc.de/
Urkund by group of teachers	2000	E	yes	single	no(paid)	http://www.urkund.com/
Copycatch by CFL Software	2002	I/E	yes	single	no(paid)	www.copycatchgold.com
Wcopyfind by Louis A. Bloomfield	2004	I/E	yes	single	no(free)	http://www.plagiarism.phys.virginia/
EVE2 by Canexus	*2001*	E	yes	single	yes(paid)	www.canexus.com
GPSP by Gllat consulting service	1999	I	yes	single	yes(paid)	http://www.plagiarism.com/
MOSS by Alex Aiken	1994	E	No	multiple	no(free)	http://theory.stanford.edu/ aiken/moss/
Jplag by Lutz Prechelt et al.	1997	E	Yes	multiple	yes(free)	https://jplag.ipd.kit.edu/
Copyscape by Indigo Stream Technologies Ltd	2011	E	yes	url	yes(free)	http://www.copyscape.com/
DOC Cop	2006	E	?	single	no(free)	www.doccop.com/
Ephorus by Ephorus B.V.	?	E	yes	single	no(paid)	http://www.ephorus.com/
iThenticate by iParadigms, LLC	1996	E	yes	single	no(paid)	http://www.ithenticate.com/
PlagiarismDetect	2008	E	yes	single	no(free)	plagiarismdetect.org/
Exactus Like by Ilya Sochenkov et al.	2016	E	yes	single	no(free)	http://like.exactus.ru/index.php/en
DupliChecker	2006	E	yes	single	no(free)	www.duplichecker.com/
Plagiarisma	?	E	yes	single	no(free)	http://plagiarisma.net/
PlagiarismChecker by Darren Hom	2006	E	yes	single	no(free)	http://www.plagiarismchecker.com/
Plagium by Septet Systems Inc.	2006	E	yes	single	no(free)	http://www.plagium.com/
PlagTracker by Svetlana et al.	2011	E	yes	single	no(paid)	http://www.plagtracker.com/
Quetext	?	I/E	yes	single	no(free)	http://www.quetext.com/
Turnitin by iParadigms	2000	E	yes	single	no(paid)	http://www.turnitin.com/
Viper by All Answers Limited	2007	E	yes	single	yes(free)	http://www.scanmyessay.com/
Maulik by Urvashi Garg et al.	2016	E		single	no	Not available
Plagiarism Scanner	2008	E	yes	single	no(paid)	http://www.plagiarismscanner.com/
Hawk Eye by Karuna Puri et al.	2015	E	yes	image		Not available
Code Match by S.A.F.E	?	E	yes	?	yes(paid)	http://www.safe-corp.com/
SID by Xin Chen et al.	2004	E	yes	single	yes(free)	http://software.bioinformatics.uwaterloo.ca/SID/.
SIM by D Gitchell	1999	E	No	multiple	yes(free)	http://www.cs.vu.nl/dick/sim.html
YAP3 by Michael J. Wise	1996	E		multiple	no	Not available
PlagScan by PlagScan GmbH	2015	E	yes	single	no(free)	www.plagscan.com/
Note: ?: Need further research						

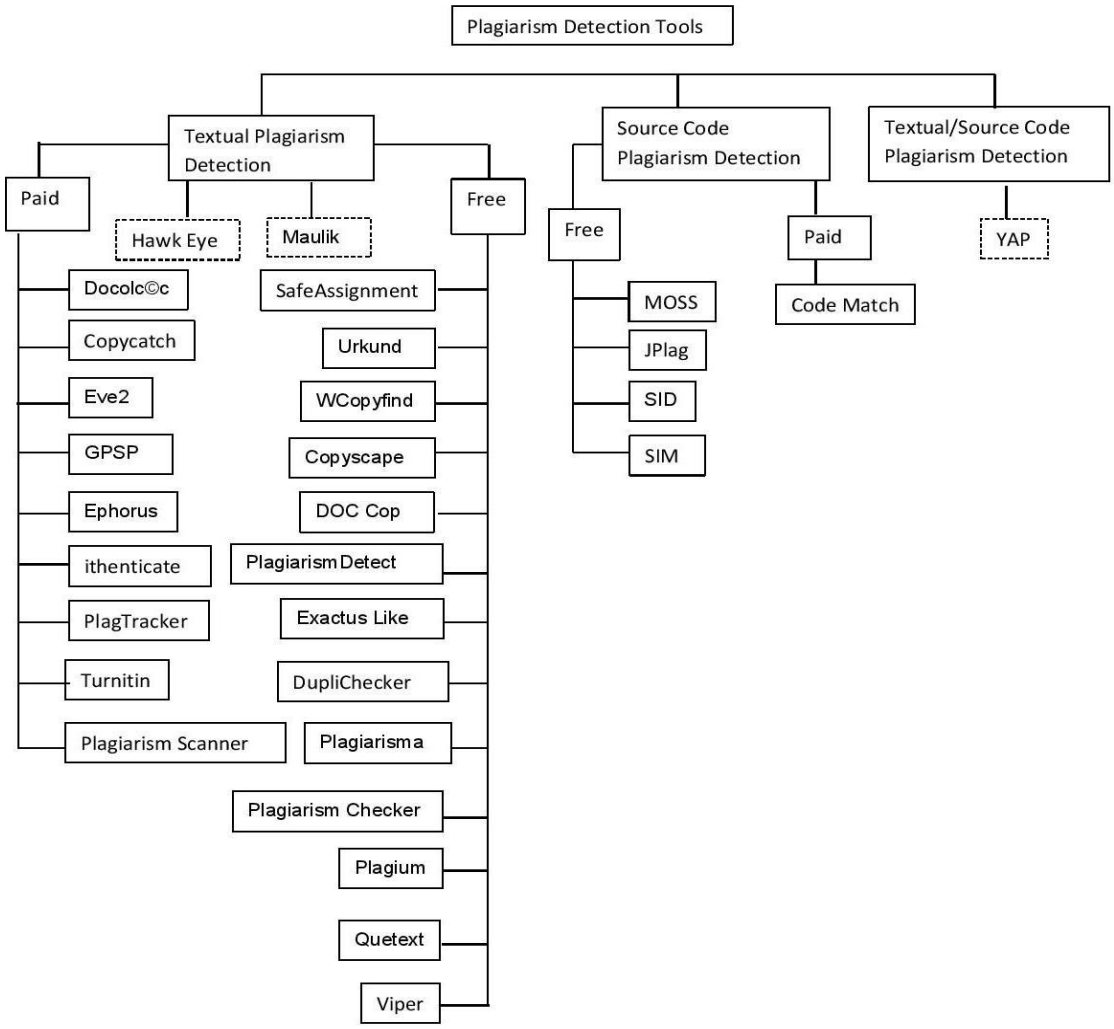

Figure 4: Classification of Plagiarism Detection Tools.

customer an opportunity to share plagiarism reports with other people by simply giving them the link, generated by this tool.

xxvi. **Hawk Eye [60]:** It is an innovative plagiarism detection system. This uses mobile scanner OCR(Optical Character Recognition) engine into convert image to text and that text it uses as input. The OCR Engine preprocess the clicked image in order to remove noise and disturbance from it and extract relevant keywords from image. The system uses plagiarism detection algorithms to remove unnecessary details like comments and changing variables names. or It uses string matching to detect plagiarism. It considers many limitations of existing well known plagiarism detection tools like Moss, JPlag, and Turnitin.

xxvii. **Code Match [8]:** Code Match compares source code and executable to detect plagiarism. It is developed by SAFE(Software Analysis and Forensic Engineering). It has also some additional functionality, which allows finding open source code within proprietary code, determining common authorship of two different programs, or discovering common, standard algorithms within different programs. It supports almost all existing programming languages.

xxviii. **SID-Software Integrity Diagnosis system [61]:** It detects plagiarism between programs by computing the shared information. It uses a metric in measuring the amount of shared information between two computer programs, to enable plagiarism detection and the metric is approximated by a heuristic compression algorithm. SID works in two phases. In the first phase, source programs are parsed to generate token sequences by a standard lexical analyser. In the second phase, Token Compress

algorithm is used to compute heuristically the shared information metric d(x; y) between each program pair within the submitted corpus. Finally, all the program pairs are ranked by their similarity distances.

xxix. **SIM [62]:** This tool is to measure similarity between two C programs. It is useful for detection of plagiarism among a large set of homework programs. This tool is robust to common modifications such as name changes, reordering of statements and functions, and adding/removing comments and white spaces.

xxx. **YAP3[63]:** YAP is a system for detecting suspected plagiarism in computer program and other text submitted by the students. YAP3 is the third version of YAP which works in two phases. In the first phase, the source text is processed to generate token sequence. In second phase, each token is (non-redundantly) compared against all others strings.

xxxi. **PlagScan [64]:** PlagScan has separate packages for schools, universities and companies. To use this we need a paid account to open. It is not a free service but if someone does not like the service, membership cancellation is possible and money will be refunded.

5. Issues and Challenges

Based on our survey we observe that in past two decades, a large number of methods and tools have been developed to support fast and accurate plagiarism detection. 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. However, with the rapid growth of digital technology to support its reproduction, storage and dissemination, some important issues and research challenges are still left unattended. In this section, we highlight some of such issues and challenges that need to be addressed by computer science and linguistic researchers.

i. 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.

ii. 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.

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

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

v. 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.

6. Conclusions

This paper has reported an exhaustive survey on plagiarism detection methods and tools in a systematic way. It has presented a taxonomy of various forms of plagiarism occur in text data and source code. Next, it has reported a large number of methods and tools under various categories and compared and analysed their pros and cons. Although in the past two decades, a large number of methods and tools have been introduced, we feel that there are still several issues and challenges left unattended. So, finally, we have highlight 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

1. S. Brin, J. Davis, H. Garcia-Molina, Copy detection mechanisms for digital documents, in: ACM SIGMOD Record, Vol. 24, ACM, 1995, pp. 398-409. 24
2. A. Parker, et al., Computer algorithms for plagiarism detection.
3. M. S. Anderson, N. H. Steneck, The problem of plagiarism, in: Urologic Oncology: Seminars and Original Investigations, Vol. 29, Elsevier, 2011, pp. 90-94.
4. N. Charya, K. Doshi, S. Bawkar, R. Shankarmani, Intrinsic plagiarism detection in digital data.
5. S. M. Alzahrani, N. Salim, A. Abraham, Understanding plagiarism linguistic patterns, textual features, and detection methods, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42 (2) (2012) 133-149.
6. H. A. Maurer, F. Kappe, B. Zaka, Plagiarism-a survey., J. UCS 12 (8) (2006) 1050-1084.
7. A. Bin-Habtoor, M. Zaher, A survey on plagiarism detection systems, International Journal of Computer Theory and Engineering 4 (2) (2012) 185.
8. R. R. Naik, M. B. Landge, C. N. Mahender, A review on plagiarism detection tools, International Journal of Computer Applications 125 (11).
9. A. M. E. T. Ali, H. M. D. Abdulla, V. Snasel, Overview and comparison of plagiarism detection tools., in: DATESO, Citeseer, 2011, pp. 161{172.
10. C. Barnbaum, Plagiarism: A student's guide to recognizing it and avoiding it.[online].[cit. 2010-12-14] (2009).
11. V. Alekya, S. S. S. Reddy, Survey of programming plagiarism detection.
12. V. Scherbinin, S. Butakov, Using microsoft sql server platform for plagiarism detection, in: Proc. SEPLN, 2009, pp. 36{37.
13. Z. Su, B.-R. Ahn, K.-Y. Eom, M.-K. Kang, J.-P. Kim, M.-K. Kim, Plagiarism detection using the levenshtein distance and smith-waterman algorithm, in: Innovative Computing Information and Control, 2008. ICICIC' 08. 3rd International Conference on, IEEE, 2008, pp. 569-569.
14. M. Elhadi, A. 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, IEEE, 2009, pp. 679--684.
15. D. R. White, M. S. Joy, Sentence-based natural language plagiarism detection, Journal on Educational Resources in Computing (JERIC) 4 (4) (2004) 2.
16. L. Moussiades, A. Vakali, Pdetect: A clustering approach for detecting plagiarism in source code datasets, The computer journal 48 (6) (2005) 651--661.
17. R. Kuppers, S. Conrad, A set-based approach to plagiarism detection., in: CLEF (Online Working Notes/Labs/Workshop), 2012.
18. A. Barron-Cedeno, P. Rosso, On automatic plagiarism detection based on n-grams comparison, in: European Conference on Information Retrieval, Springer, 2009, pp. 696-700.
19. T. C. Hoad, J. Zobel, Methods for identifying versioned and plagiarized documents, Journal of the American society for information science and technology 54 (3) (2003) 203-215.
20. C. Grozea, C. Gehl, M. Popescu, Encoplot: Pairwise sequence matching in linear time applied to plagiarism detection, in: 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse, 2009, p. 10.
21. C. Basile, D. Benedetto, E. Caglioti, G. Cristadoro, M. Esposti, A plagiarism detection procedure in three steps: Selection, matches and squares, in: Proc. SEPLN, 2009, pp. 19-23.
22. H. Zhang, T. W. Chow, A coarse-to- ne framework to efficiently thwart plagiarism, Pattern Recognition 44 (2) (2011) 471-487.

23. M. Elhadi, A. Al-Tobi, Use of text syntactical structures in detection of document duplicates, in: Digital Information Management, 2008. ICDIM 2008. Third International Conference on, IEEE, 2008, pp. 520-525.
24. C. Fellbaum, WordNet, Wiley Online Library, 1998.
25. S. Torres, A. Gelbukh, Comparing similarity measures for original wsd lesk algorithm, Research in Computing Science 43 (2009) 155-166.
26. P. 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 (1999) 95--130.
27. C. Leacock, G. A. Miller, M. Chodorow, Using corpus statistics and wordnet relations for sense identification, Computational Linguistics 24 (1) (1998) 147-165.
28. R. Yerra, Y.-K. Ng, A sentence-based copy detection approach for web documents, in: International Conference on Fuzzy Systems and Knowledge Discovery, Springer, 2005, pp. 557--570.
29. J. Koberstein, Y.-K. Ng, Using word clusters to detect similar web documents, in: International Conference on Knowledge Science, Engineering and Management, Springer, 2006, pp. 215--228.
30. S. M. Alzahrani, N. Salim, On the use of fuzzy information retrieval for gauging similarity of arabic documents, in: Applications of Digital Information and Web Technologies, 2009. ICADIWT'09. Second International Conference on the, IEEE, 2009, pp. 539-544.
31. M. Rahman, W. P. Yang, T. W. Chow, S. Wu, A exible multi-layer self-organizing map for generic processing of tree-structured data, Pattern Recognition 40 (5) (2007) 1406--1424.
32. T. W. Chow, M. Rahman, Multilayer som with tree-structured data for efficient document retrieval and plagiarism detection, IEEE Transactions on Neural Networks 20 (9) (2009) 1385--1402.
33. S. M. Zu Eissen, B. Stein, M. Kulig, Plagiarism detection without reference collections, in: Advances in data analysis, Springer, 2007, pp. 359--366.
34. E. Stamatatos, A survey of modern authorship attribution methods, Journal of the American Society for information Science and Technology 60 (3) (2009) 538-556.
35. M. Potthast, A. Barron-Cede˜no, B. Stein, P. Rosso, Cross-language plagiarism detection, Language Resources and Evaluation 45 (1) (2011) 45--62.
36. A. H. Osman, N. Salim, A. Abuobieda, Survey of text plagiarism detection, Computer Engineering and Applications Journal (ComEngApp) 1 (1) (2012) 37-45.
37. J.-P. Bao, J.-Y. Shen, X.-D. Liu, Q.-B. Song, A survey on natural language text copy detection, Journal of software 14 (10) (2003) 1753{1760.
38. M. S. Binwahlan, N. Salim, L. Suanmali, Fuzzy swarm diversity hybrid model for text summarization, Information processing & management 46 (5) (2010) 571-588.
39. V. Mitra, C.-J. Wang, S. Banerjee, Text classification: A least square support vector machine approach, Applied Soft Computing 7 (3) (2007) 908-914.
40. D. Zou, W.-J. Long, Z. Ling, A cluster-based plagiarism detection method, in: Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September, 2010.
41. M. Zini, M. Fabbri, M. Moneglia, A. 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), IEEE, 2006, pp. 181--185.
42. A. Si, H. V. Leong, R. W. Lau, Check: a document plagiarism detection system, in: Proceedings of the 1997 ACM symposium on Applied computing, ACM, 1997, pp. 70-77.
43. B. Gipp, J. 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, ACM, 2010, pp. 273--274.
44. B. Gipp, N. 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, ACM, 2011, pp. 249--258.
45. D. Atkinson, S. Yeoh, Student and sta perceptions of the effectiveness of plagiarism detection software, Australasian Journal of Educational Technology 24 (2) (2008) 222-240.
46. R. A. Ahmed, Overview of different plagiarism detection tools.
47. L. Prechelt, G. Malpohl, M. Philippsen, Finding plagiarisms among a set of programs with jplag, J. UCS 8 (11) (2002) 1016.
48. B. Scaife, Plagiarism detection software report, NCC Group.
49. E. A. Ochroch, Review of plagiarism detection freeware, Anesthesia & Analgesia 112 (3) (2011) 742--743.
50. U. Garg, Plagiarism and detection tools: An overview, Research Cell: An International Journal of Engineering Sciences 2 (2011) 92--97.
51. http://www.ithenticate.com/, ithenticate.
52. I. Sochenkov, D. Zubarev, I. Tikhomirov, I. Smirnov, A. Shelmanov, R. Suvorov, G. Osipov, Exactus like: Plagiarism detection in scientific texts, in: European Conference on Information Retrieval, Springer, 2016, pp. 837--840.
53. http://www.duplichecker.com/, Duplichecker.
54. http://plagiarisma.net/, Plagiarisma.
55. http://www.plagium.com/en/plagiarismchecker, Plagiarismchecker.
56. http://www.plagtracker.com/, Plagtracker.
57. http://www.quetext.com/, Quetext.
58. A. H. Osman, N. Salim, M. S. Binwahlan, Plagiarism detection using graphbased representation, arXiv preprint arXiv:1004.4449.
59. U. Garg, V. Goyal, Maulik: A plagiarism detection tool for hindi documents, Indian Journal of Science and Technology 9 (12).
60. P. Mulay, K. Puri, Hawk eye: Intelligent analysis of socio inspired cohorts for plagiarism, in: Innovations in Bio-Inspired Computing and Applications, Springer, 2016, pp. 29--42.
61. X. Chen, B. Francia, M. Li, B. Mckinnon, A. Seker, Shared information and program plagiarism detection, IEEE Transactions on Information Theory 50 (7) (2004) 1545-1551.
62. D. Gitchell, N. Tran, Sim: a utility for detecting similarity in computer programs, in: ACM SIGCSE Bulletin, Vol. 31, ACM, 1999, pp. 266--270.
63. M. J. Wise, Yap3: Improved detection of similarities in computer program and other texts, ACM SIGCSE Bulletin 28 (1) (1996) 130--134.
64. http://www.plagscan.com/, Plagscan.

Plagiarism: A Plague

Richa Tripathi **S Kumar**

Abstract

This paper defines the term ' plagiarism', and discusses about the tools, types, studies on plagiarism and the methods to control plagiarism. It also suggests for submission of E-thesis and preparation of databases of thesis and projects in India.

Keywords: Plagiarism, Anti-Plagiarism, JISC

1. Introduction

Plagiarism is an issue of great concern amongst the academicians. Plagiarism is a moral, ethical, and legal issue. Plagiarism has been around for centuries, but the Internet and the subsequent proliferation of information have made the problem more serious. Plagiarism is taking someone else's work and passing it off as one's own. Many people think of plagiarism as copying another's work, or borrowing someone else's original ideas. But terms like "copying" and "borrowing" can disguise the seriousness of the offense. Dictionary definition of this tem is as follows:

1. To steal and pass off (the ideas or words of another) as one's own
2. To use (another's production) without crediting the source
3. To commit literary theft
4. To present as new and original an idea or product derived from an existing source.
5. In other words, plagiarism is an act of fraud. It involves both stealing someone else's work and lying about it afterward.

2. Types of Plagiarism

Plagiarism includes copying words or ideas from someone else without giving credit; failing to put a quotation in quotation marks; giving incorrect information about the source of a quotation; changing words but copying the sentence structure of a source without giving credit; copying so many words or ideas from a source that it makes up the majority of your work . The types of Plagiarism can be categorized and listed as given below:

2.1 Sources Not Cited

2.1.1 The Ghost Writer

The writer turns in another's work, word-for-word, as his or her own.

2.1.2 The Photocopy

The writer copies significant portions of text straight from a single source, without alteration.

2.1.3 The Potluck Paper

The writer tries to disguise plagiarism by copying from several different sources, tweaking the

sentences to make them fit together while etaining most of the original phrasing.

2.1.4 The Poor Disguise

Although the writer has retained the essential content of the source, he or she has altered the paper's appearance slightly by changing key words and phrases.

2.1.5 The Labor of Laziness

The writer takes the time to paraphrase most of the paper from other sources and make it all fit together, instead of spending the same effort on original work.

2.1.6 The Self-Stealer

The writer "borrows" generously from his or her previous work, violating policies concerning the expectation of originality adopted by most academic institutions.

2.2 Sources Cited (But Still Plagiarized)

2.2.1 The Forgotten Footnote

The writer mentions an author's name for a source, but neglects to include specific information on the location of the material referenced. This often masks other forms of plagiarism by obscuring source locations.

2.2.2 Misinformed

The writer provides inaccurate information regarding the sources, making it impossible to find them.

2.2.3 The Too-Perfect Paraphrase

The writer properly cites a source, but neglects to put in quotation marks text that has been copied word-for-word, or close to it. Although attributing the basic ideas to the source, the writer is falsely claiming original presentation and interpretation of the information.

2.2.4 The Resourceful Citer

The writer properly cites all sources, paraphrasing and using quotations appropriately. The paper contains almost no original work! It is sometimes difficult to spot this form of plagiarism because it looks like any other well-researched document.

2.2.5 The Perfect Crime

Well, we all know it doesn't exist. In this case, the writer properly quotes and cites sources in some places, but goes on to paraphrase other arguments from those sources without citation. This way, the writer tries to pass off the paraphrased material as his or her own analysis of the cited material.

2.3 Other Types of Plagiarism

Other types of plagiarism have also been recognized. These are:

2.3.1 Copy and Paste Plagiarism

Any time a sentence or significant phrase intact from a source is lifted , you must use quotation marks and reference the source.

2.3.2 Word Switch Plagiarism

If you take a sentence from a source and change around a few words, it is still plagiarism. If you want to quote a sentence, then you need to put it in quotation marks and cite the author and article. But quoting Source articles should only be done if what the quote says is particularly useful in the point you are trying to make in what you are writing. In many cases, a quotation would not really be useful. The

person who plagiarizes is sometimes just too lazy to synthesize the ideas expressed in the Source article.

2.3.3 Metaphor Plagiarism

Metaphors are used either to make an idea clearer or give the reader an analogy that touches the senses or emotions better than a plain description of the object or process. Metaphors, then, are an important part of an author's creative style. If you cannot come up with your own metaphor to illustrate an important idea, then use the metaphor in the Source Article, but give the author credit for it.

2.3.4 Idea Plagiarism

If the author of the source article expresses a creative idea or suggests a solution to a problem, the idea or solution must be clearly attributed to the author.

2.3.5 Reasoning Style/Organization Plagiarism

When you follow a Source Article sentence-by-sentence or paragraph-by-paragraph, it is plagiarism, even though none of your sentences are exactly like those in the Source article or even in the same order. What you are copying in this case is the author's reasoning style.

2.4 Data Plagiarism

In research, often data is plagiarized.

3. Anti-Plagiarism Tools

Where there is an ailment, there is a treatment. So is true with plagiarism and there are many anti-plagiarism tools.

3.1 CopyCatch Gold

http://www.copycatch.freeserve.co.uk/

A forensic linguist at CFL Software Development with extensive experience in plagiarism developed this software for teachers and students. The cost of a single user license for educational use is £250 per year.

3.2 EduTie.com

http://www.edutie.com/

EduTie.com was founded in August 2000, and is designed to help institutions prevent Internet plagiarism. It is built on the PlagiServe (http://www.plagiserve.com) core design. Papers submitted are compared to more than 1 billion "high risk" Web pages in an attempt to detect plagiarism. Free trials of the software are available.

3.3 EVE2: Essay Verification Engine

http://www.canexus.com/eve/index.shtml

EVE2 claims to come as close as possible to searching every site on the Internet to detect plagiarism by "employing the most advanced searching tools available to locate suspected sites. Free fifteen day trials are available, but the software must be purchased after that time to continue using it. Each license is a one-time fee of $19.99 and updates are free.

3.4 Glatt Plagiarism Program

http://www.plagiarism.com

Dr. Barbara Glatt has developed the 3 different software programs designed to detect and prevent plagiarism. The 3 parts are the Plagiarism Teaching Program, the Plagiarism Screening Program and the Plagiarism Self-Detection Program. Costs for the programs runs around $250 each if bought as a complete set or $300 if purchased individually.

A list of publications that have reviewed the Glatt Plagiarism Program can be found. at http://www.plagiarism.com/publications.htm.

3.5 Google

http://www.google.com

Google is not designed to be a plagiarism detection tool, but its advanced search engine capabilities are conducive to locating key phrases that may appear in students' research papers. The Google Directory also has numerous links to information about plagiarism detection devices at http://directory.google.com/Top/Reference/Education/Educators/Plagiarism/Detection/.

3.6 Joint Information Systems Committee (JISC): Electronic Plagiarism Detection http://www.jisc.ac.uk/plagiarism/

JISC completed a plagiarism project in 2001, and they are establishing a plagiarism advisory service as a result of this experience. There were 4 parts to their plagiarism project, and they include:

1. Technical review of free-text plagiarism detection software
2. Technical review of source code plagiarism detection software
3. A pilot of free-text detection software in 5 UK institutions
4. A good practice guide to plagiarism detection

A listserv has also been established to continue discussions dealing with academic dishonest and plagiarism issues.

A copy of JISC's Technical Review of Plagiarism Detection Software Report can be accessed at http://www.jisc.ac.uk/pub01/luton.pdf.

3.7 JISC Plagiarism Advisory Service http://online.northumbria.ac.uk/faculties/art/information_studies/Imri/JISCPAS/site/default.htm

JISC Plagiarism Advisory Service is a new offering that began in September 2002. It is based in the Information Management Research Institute at Northumbria University (UK). New materials are constantly being added to this plagiarism portal, but it currently offers advice & guidance, educational materials for students and other online resources. A plagiarism detection service, supported by the Joint Information Systems Committee (JISC) until August 2004, is based on the turnitin.com platform and allows instructors to conduct electronic comparisons of work complete by students.

3.8 Jplag

http://www.jplag.de/

Guido Malpohl initially developed this software which is designed to detect academic dishonesty. The software does more than merely compare the text of documents. JPlag also looks at program language syntax and program structure so it can also be used to detect stolen software parts. Instructors may use JPlag for free, but they must first set up an account in order to prevent unauthorized use by students.

3.9 Library Electronic Databases

http://gateway.library.uiuc.edu/ersearch/

The Library at the University of Illinois at Urbana-Champaign provides access to numerous electronic resources for students and faculty. Instructors may want to consult these resources when checking for plagiarism.

Full text databases like EBSCO and Expanded Academic ASAP (InfoTrac) are two obvious starting points when checking undergraduate assignments. One thing to keep in mind is that some resources that are not full text but provide abstract information are often used by students.

3.10 MOSS

http://www.cs.berkeley.edu/~aiken/moss.html

Moss or Measure of Software Similarity is a tool that has been used primarily to detect plagiarism. The way it works is that it detects similarities of C, C++, Java, Pascal, Ada, ML, Lisp or Scheme programs. Moss is free to use for instructors and staff of programming language courses only.

3.11 Plagiarism.org

http://www.plagiarism.org

University of California Berkeley students and alumni created plagiarism.org to be used to detect plagiarism. One thing to watch out for is that the software doesn't differentiate between quoted materials and original writing.

3.12 The Plagiarism Resource Site

http://www.plagiarism.phys.virginia.edu/

Lou Bloomfield, Professor of Physics at the University of Virginia, is the sole author of The Plagiarism Resource Site. The goal of this site is to "help reduce the impact of plagiarism on education and educational institutions". Numerous links are provided to sources on how to deal with plagiarism.

3.13 PlagiServe

http://www.plagiserve.com/

Olexiy Shevchenko, Max Litvin and Sasha Lugovskyy, the PlagiServe Team, came up with the concept of a plagiarism detection device in June 2000. The software used by PlagiServe not only detects papers that have been obtained from a term paper company and turned into an instructor, but it also looks for any changes or modifications made to these papers. PlagiServe has a database of over 150,000 student essays, term papers and cliff notes, and they also send out Web robots to check "high risk" sites like Britannica.com, Refdesk.com and Encyclopedia.com for copied materials. NOTE: Instructors may want to be careful about using this particular detection device. Some indicate it may also sell term papers to students.

3.14 Turnitin

http://www.turnitin.com/

Turnitin, a plagiarism.org partner, considers themselves to be "the world's most widely recognized and trusted resource to prevent Internet plagiarism". Free trials are also available, and subscription costs vary depending on the type of plan chosen.

Turnitin is currently the subject of a copyright controversy. For more information, check out the following article, "A Plagiarism Detection Tool Creates Legal Quandary" at http://chronicle.com/free/v48/i36/36a03701.htm.

3.15 WordCHECK

http://www.wordchecksystems.com/

WordCHECK is used by a diverse group including information researchers, copyright attorneys and classroom teachers. This plagiarism detection device was developed by Information Analytics, a Lincoln, NE company owned by Kenneth Livingston and Mark Dahmke. WordCHECK may be purchased for a fee.

4. Suggestions

The following suggestions are made to avoid plagiarism:

- Preparation of data bases of thesis for Ph.D submitted to universities.

- Compulsory submission of Electronic copy in a data base of the UGC which should be open before award so that any one can detect plagiarism & bring to the universities.
- Prepration of data bases of articles published in conferences & journals in India which are not covered international database.
- Taking an affidavit from the candidate regarding no use of plagirised material.

5. Conclusion

In India, in absence of database of theses and dissertations, it is easy to plagiarize them from one university to anther and even in the same university. Plagiarism in project work is common and difficult to detect. I come across unique form of plagiarism where photocopies M.L.I.Sc dissertation have been used by another college in the different years. Also in the case in many other subjects Ph.D. thesis are also alleged to use plagiarism practices. In another kind, Ph.D. thesis includes word toward translation from English to Hindi of a book in most of its chapter. India requires an anti-plagiarism policy in academic and develop its own anti plagiarism software. With the above discussion it is clear that if plagiarism is easy the tools to detect the plagiarism are also available easily. Beware and avoid plagiarism otherwise you may be another example being quoted "XYZ Controversy."

References

1. http://www.plagrism.org
2. http://en.wikipedia.org/wiki/scientific-plagiarism-in-india
3. http://www.unmc.edu/library/plagrism
4. http://www.usp.edu/writing/plagrism.shtml
5. http://terpconnect.umd.edu/~toh/reseach/
6. http://www.indiana.edu/~wts/pamphlets/plagiarism.shtmltears
7. http://www.uow.edu.au/handbook/courserules/plagiarism.html
8. http://www.geneseo.edu/~brainard/plagiarismtypes.htm
9. http://my.fit.edu/~jbarlow/HBB/plagiriasmstuff/plagiarismdetectiontools.htm
10. http://www.expresscomputeronline.com/20080310/technology06.shtml
11- http://www.shamples.net/pages/staff/ptools/
12. http://www.123helpme.com/plagiarism.jsp
13. http://www.india.edu/~wts/pamphlets/plagiarism.shtml
14. http://www.unc.edu/depts./wcweb/handouts/plagiarism.html
15. http://www.mantex.co.uk/samples/plagiarism.htm
16. http:/www.answer.com/topic/plagiarism

About Authors

Ms. Richa Tripathi, Resarch Scholar, Vikram University, Ujjain.

Prof. S Kumar, Professor, Vikram University, Ujjain.

Overview and Comparison of Plagiarism Detection Tools

Asim M. El Tahir Ali, Hussam M. Dahwa Abdulla, and Václav Snášel

Department of Computer Science, VŠB-Technical University of Ostrava,
17. listopadu 15, Ostrava - Poruba, Czech Republic
asim070@yahoo.com, hussamdahwa@hotmail.com, vaclav.snasel@vsb.cz

Abstract. In this paper we have done an overview of effective plagiarism detection methods that have been used for natural language text plagiarism detection, external plagiarism detection, clustering-base plagiarism detection and some methods used in code source plagiarism detection, also we have done a comparison between five of software used for textual plagiarism detection: (PlagAware, PlagScan, Check for Plagiarism, iThenticate and PlagiarismDetection.org), software are compared with respect of their features and performance.

1 Introduction

"Plagiarism, the act of taking the writings of another person and passing them off as one's own. The fraudulence is closely related to forgery and piracy-practices generally in violation of copyright laws." Encyclopedia Britannica [5].

Plagiarism can be considered as one of the electronic crimes, like (computer hacking, computer viruses, spamming, phishing, copyrights violation and others crimes). Plagiarism defined as the act of taking or attempting to take or to use (whole or parts) of another person's works, without referencing or citation him as the owner of this work. It may include direct copy and paste, modification or changing some words of the original information from the internet books, magazine, newspaper, research, journal, personal information or ideas.
According to the Merriam-Webster Online Dictionary, to "plagiarize" means:

- To steal and pass off (the ideas or words of another) as one's own.
- To use (another's production) without crediting the source.
- To commit literary theft.
- To present as new and original an idea or product derived from an existing source.

Also according to Turnitin.com, plagiarism.org and Research Resources this are considered plagiarism:

- Turning in someone else's work as your own.
- Copying words or ideas from someone else without giving credit.
- Failing to put a quotation in quotation marks.

V. Snášel, J. Pokorný, K. Richta (Eds.): Dateso 2011, pp. 161–172, ISBN 978-80-248-2391-1.

- Giving incorrect information about the source of a quotation.
- Changing words but copying the sentence structure of a source without giving credit.
- Copying so many words or ideas from a source that it makes up the majority of your work, whether you give credit or not (see our section on "fair use" rules).

Plagiarism can be classified into five categories:

1. Copy & Paste Plagiarism.
2. Word Switch Plagiarism.
3. Style Plagiarism.
4. Metaphor Plagiarism.
5. Idea Plagiarism.

There are two types of plagiarism are more occurs:

1. Textual plagiarisms: this type of plagiarism usually done by students or researchers in academic enterprises, where documents are identical or typical to the original documents, reports, essays scientific papers and art design.
2. A source code plagiarism: also done by students in universities, where the students trying or copying the whole or the parts of source code written by someone else as one's own, this types of plagiarism it is difficult to detect.

2 Why Plagiarism Detection is Important

In some of the academic enterprises like universities, schools and institutions, plagiarism detection and prevention became one of the educational challenges, because most of the students or researchers are cheating when they do the assigned tasks and projects. This is because a lot of resources can be found on the internet. It is so easy to them to use one of the search engines to search for any topic and to cheat from it without citing the owner of the document. So it is better and must all academic fields they should have to use plagiarism detection soft-wares to stop or to eliminate students cheating, copying and modifying documents when they know that they will be found.

Some types of plagiarism acts can be detected easily by using some of the recent plagiarism detection soft-wares available on the market or over the internet. However for some of the expert plagiarism who is using some of the anti-plagiarism soft-wares which are available over the internet, it needs more efforts to detect the plagiarism or cannot be detected at all.

Plagiarism is practiced not only by student but also there are some staff members who like to publish papers in which some parts are directly copied or partially modified to be one of the famous people.

There is a big number of plagiarism soft-wares used for plagiarism detection and many of detection tools have been developed by researchers but still they have some limitations as they cannot prove or they show evidence that the documents has been plagiarized from another document or sources it only shows

the similarity and give hints to some other documents. This is if the paper has been published globally in some international journal, but some of universities and some of the research centers still do not taking any action against plagiarism detection which help people to cheat more and more.

So still now by using the recent detection software, plagiarism can not 100% be detected?

Copyrights and legal aspects for use of published documents also can be covered by using plagiarism software, so it can show whether this person has legally or illegally copied the documents or not and it also show the whether this person has permission from the owner to use this document or not.

Plagiarism detection is also one of the most important issues to journals, research center and conferences; they are using advanced plagiarism detection tools to ensure that all the documents have not been plagiarized, and to save the copyrights from violation for the publishers.

3 Plagiarism Detection Methods

In both the textual document plagiarism and source code plagiarism, detection can be either: Manual detection or automatic detection.

- Manual detection: done manually by human, its suitable for lectures and teachers in checking student's assignments but it is not effective and impractical for a large number of documents and not economical also need highly effort and wasting time.
- Automatic detection (Computer assisted detection): there are many software and tools used in automatic plagiarism detection, like PlagAware, PlagScan, Check for Plagiarism, iThenticate, PlagiarismDetection.org, Academic Plagiarism, The Plagiarism Checker, Urkund, Docoloc and more.

3.1 Textual Plagiarism

Many of researchers are developed a set of tools used in textual automatic detection like:

Grammar-based method The grammar-based method is important tool to detect plagiarism. It focuses on the grammatical structure of documents, and this method uses a string-based matching approach to detect and to measure similarity between the documents. The grammar- based methods is suitable for detecting exact copy without any modification, but it is not suitable for detecting modified copied text by rewriting or switching some words that has the same meaning. This is considered as one of this method limitations [4].

Semantics-based method The semantics-based method, also considered as one of the important method for plagiarism detection, focuses on detecting the similarities between documents by using the vector space model. It also can calculate and count the redundancy of the word in the document, and then they use the fingerprints for each document for matching it with fingerprints in other documents and find out the similarity. The semantic-based method is suitable for non partial plagiarism as mentioned before use the whole document and use vector space to match between the documents, but if the document has been partially plagiarized it cannot achieve good results, and this is considered as one of the limitations of this method, because it is difficult to fix the place of copied text in the original document [4, 1].

Grammar semantics hybrid method Grammar semantic hybrid method is considered as the most important method in plagiarism detecting for the natural languages. This method, so effective in achieving better and improving plagiarism detection result, is suitable for the copied text including modified text by rewriting or switching some words that have the same meaning, which cannot be detected by grammar-based method. It also solves the limitation of semantic-based method. Grammar semantic hybrid method can detect and determine the location of plagiarized parts of the document, which cannot be detected by semantic-based method, and calculating the similarity between documents [4, 1].

External plagiarism detection method The external plagiarism detection relies on a reference corpus composed of documents from which passages might have been plagiarized A passage could be made up of paragraphs, a fixed size block of words, a block of sentences and so on. A suspicious document is checked for plagiarism by searching for passages that are duplicates or near duplicates of passages in documents within the reference corpus. An external plagiarism system then reports these findings to a human controller who decides whether the detected passages are plagiarized or not. A naive solution to this problem is to compare each passage in a suspicious document to every passage of each document in the reference corpus. This is obviously prohibitive. The reference corpus has to be large in order to find as many plagiarized passages as possible [20].

This fact directly translates to very high runtimes when using the naive approach. External plagiarism detection is similar to textual information retrieval (IR) [3]. Given a set of query terms an IR system returns a ranked set of documents from a corpus that best matches the query terms. The most common structure for answering such queries is an inverted index. An external plagiarism detection system using an inverted index indexes passage of the reference corpus' documents.

Such a system was presented in [7] for finding duplicate or near duplicate documents.

Another method for finding duplicates and near duplicates is based on hashing or fingerprinting. Such methods produce one or more fingerprints that de-

scribe the content of a document or passage. A suspicious document's passages are compared to the reference corpus based on their hashes or fingerprints. Duplicate and near duplicate passages are assumed to have similar fingerprints. One of the first systems for plagiarism detection using this schema was presented in [2]. External plagiarism detection can also be viewed as nearest neighbor problem in a vector space R^d.

Clustering in plagiarism detection Document clustering is one of the important techniques used by information retrieval in many purposes; it has been used in summarization of the documents to improve the retrieval of data by reducing the searching time in locating the document. It is also used for result presentation. Document clustering is used in plagiarism detection to reduce the searching time. But still now in clustering there are some limitations and difficulties with time and space [8].

Most of the above methods have been used by textual documents plagiarism detection.

3.2 Source code plagiarism

Source code plagiarism or it called programming plagiarisms usually done by students in universities and schools can be defined act or trial to use, reuse, convert and modify or copy the whole or the part of the source code written by someone else and used in your programming without citation to the owners. Source code detection mainly requires human intervention if they use Manual or automatic source code plagiarism detection to decide or to determine whether the similarity due to the plagiarism or not. Manual detection of source code in a big number of student homework's or project it is so difficult and needs highly effort and stronger memory, it seems that impossible for a big number of sources.

Plagiarism detection system or algorithms used in source-code similarity detection can be classifies according to Roy and Cordy [9] can be classified as based on either:

- 'Strings - look for exact textual matches of segments, for instance five-word runs. Fast, but can be confused by renaming identifiers'.
- "Tokens - as with strings, but using a lexer to convert the program into tokens first. This discards whitespace, comments, and identifier names, making the system more robust to simple text replacements. Most academic plagiarism detection systems work at this level, using different algorithms to measure the similarity between token sequences".
- "Parse Trees - build and compare parse trees. This allows higher-level similarities to be detected. For instance, tree comparison can normalize conditional statements, and detect equivalent constructs as similar to each other".
- "Program Dependency Graphs (PDGs) - a PDG captures the actual flow of control in a program, and allows much higher-level equivalences to be located, at a greater expense in complexity and calculation time".

- "Metrics - metrics capture 'scores' of code segments according to certain criteria; for instance, "the number of loops and conditionals", or "the number of different variables used". Metrics are simple to calculate and can be compared quickly, but can also lead to false positives: two fragments with the same scores on a set of metrics may do entirely different things ".
- "Hybrid approaches - for instance, parse trees + suffix trees can combine the detection capability of parse trees with the speed afforded by suffix trees, a type of string-matching data structure".

There are many methods developed by researcher for source code plagiarism detection like:

- Cynthia Kustanto and Inggriani Liem: they developed a tool for automatic source code detection call Deimos, used in source plagiarism detection, to provide a clear readable form and to erase the displayed result. It was develop to be used with LISP and Pascal programming languages. The time consumed by this tool for section a number of 100 LISP was efficient [11].
- Boris Lesner, Romain Brixtel, Cyril Bazin and Guillaume Bagan: they introduced a new frame work named A Novel Framework to Detect Source Code Plagiarism, mainly used in detection of four type of code source plagiarisms which are change the code name, rebuilt or recoded again, move, add, change and remove the code and replace some text from place to place with the code. A bottom-up approach has been implemented to six steps which are: 1- first the Pre-flattering the source code: they use common method in filtering a source code that by indicating and rename each alphanumerical string in the code. 2- Second they segment the source code to segmentation and measure the similarity on it 3- thirdly they matched each segment and reposted it for filtering. 4-5: Fourthly the use matrix M that have been used in filtering in evaluation of the document 6- In this stage start to analysis the original document according to the evaluation done by document wise distance. This method has been applied to copra languages and shows a great result [12].
- Ameera Jadalla and Ashraf El Nagar: They develop Plagiarism Detection Engine used for detection of source code plagiarism for Java (PDE4Java). The proposed search engine divided in to three steps 1- step one is the process of the tokenization for the Java code 2- second step is to find and measure the similarity between the original code and the tokenized code 3- lastly is to cluster the Java code in order to be used in plagiarism detection as reference. This search engine can be used with all programming language due to its flexibility. Report can show for each cluster code besides the textual [10].

4 Comparisons

We compare the plagiarism software used in textual and source code plagiarism into two categories: first according to features and secondly according to performance [6]. Qualitative comparison used in comparing the features of software, where we are looking for properties of the tools. Quantitative comparisons used

in comparing the performance of software, where it depends on the result. Here comparison of some textual soft-wares:

4.1 PlagAware

Is an online-service used for textual plagiarism detection, which allows and offers some services for the user for example can search, find, analyze and trace plagiarism in the specified topic similar to the topics, PlagAware is a search engine, which is considered as the main element, which is strong in detecting typical contents of given texts. It uses the classical search engine for detecting and scanning plagiarism, and provide different types of report that help the user or the document owner to decide that is his document has been plagiarized or not. The two primary fields for PlagAware plagiarism search engine is webpage monitoring for theft contents and transmitted text assessment. In [13], there are three application fields of PlagAware [14]:

- Tracing content theft: Webmaster can use PlagAware for detecting and tracing plagiarisms of websites, in order to find out the plagiarized or the copied contents. PlagAware is considered as strong total solution software systems, which allow the operators of websites to do an automatic observation of their own pages against possible content theft.
- PlagAware is used in search for plagiarisms of student's academic documents and analyze them. Also it is used to assess plagiarism, also to follow and prove the origin of the works including all of academic documents. It generates report that helps them to fast detection of plagiarism.
- Proof of authorship is also provided by PlagAware: it became more important to the authors to ensure that the authorization have been granted to their publication including all types of publication this gives them additional competitive advantage and increase the value of your work.

The main features of PlagAware are [15]:

- Database Checking: PlagAware is a search engine that allows the user to submit his document and Plagaware start searching over the internet. So mainly it does not have local database but it offers checking other database that are available over the internet.
- Internet Checking: PlagAware is an online application and it considered as one of search engine, allows the student or webmaster to upload and check their academic documents, homework, manuscript and articles to be searched against plagiarism over world wide web.ans also provides a webmaster to have capability to do automatic observation of their own page against possible contents theft.
- Publications Checking: PlagAware: support mainly used in academic filed so it provides checking of most types of submitted publication like homework, manuscript, documents, including, books, articles, magazines, journals, editorial and PDFs etc.

- Synonym and Sentence Structure Checking: PlagAware does not support synonym and sentence structure checking.
- Multiple Document Comparison: PlagAware offers comparison of multiple documents.
- Supported Languages: PlagAware supports German as primary language, English and Japanese as secondary languages.

4.2 PlagScan

PlagScan is online software used for textual plagiarism checker. PlagScan is often used by school and provides different types of account with different features. PlagScan use complex algorithms for checking and analyzing uploaded document for plagiarism detection, based on up-to-date linguistic research. Unique signature extracted from the document's structure that is then compared with PlagScan database and millions of online documents. So PlagScan is able to detect most of plagiarism types either directs copy and paste or words switching, which provides an accurate measurement of the level of plagiarized content in any given documents [16]. The Main features of PlagScan are [15]:

- Database Checking: PlagScan it has own database that include millions documents like (paper, articles and assignments), and articles over World Wide Web. So it offers database checking whether locally or others database over the internet.
- Internet checking: PlagScan is an online checker so it provides internet checking to all submitted documents. Whether that the document available on the internet or available in the local database or cached.
- Publications Checking: PlagScan: is mainly used in academic filed so it provides checking most types of submitted publication like documents, including, books, articles, magazines, journals, newspapers, PDFs etc. online only.
- Synonym and Sentence Structure Checking: PlagScan does not support synonym and sentence structure checking but provides Integration via application programming interface in your existing content management system or learning management system possible.
- Multiple Document Comparison: CheckForPlagiarism.net offers comparison of multiple documents in parallel.
- Supported Languages: PlagScan supports all the language that use the international UTF-8 encoding and all language with Latin or Arabic characters can be checked for plagiarism.

4.3 CheckForPlagiarism.net

CheckForPlagiarism.net was developed by a team of professional academic people and became one of the best online plagiarism checkers that used to stop or prevention of online plagiarism and minimizes its effects on academic integrity. In order to maximize the accuracy CheckForPlagiarism.net has used the some

methods like document fingerprint and document source analysis to protect document against plagiarism. The fingerprint-based approach used to analyze and summarize collection of document and create a kind of fingerprint for it. Some of numerical attributes can be used by fingerprint that somehow reflects in the structure of the document. So by creating fingerprint for each document with some of numerical attributes for each document in the collection, we can easily find the matching or the similarity between documents across billions of articles. Using this feature by CheckForPlagiarism.net increased the efficiency in detecting most types of plagiarisms [17]. The main features of CheckForPlagiarism.net are [15]:

- Database Checking: CheckForPlagiarism.ne uses its own database that include millions documents like (paper, articles and assignments), and articles over World Wide Web. So it offers fast and reliable depth database checking, also provides checking through all other databases in different fields like medical database, law- related database and other specialty and generalized databases.
- Internet Checking: CheckForPlagiarism.ne: live(online) and cached links to websites used for extensive internet checking to all submitted documents. One more advantage is that it can still check your documents against if a website that is no longer online, this include all contents of website like forums, message boards, bulletin boards, blogs, and PDFs etc., all this check is done automatically and in (almost) real-time.
- Publications Checking: CheckForPlagiarism.net offers detailed and deep checking of most types of submitted publication documents, including, books, articles, magazines, journals, newspapers, PDFs etc. this is done whether the publications is available online (active on the internet) or not available on the internet offline (store paper based).
- Synonym & Sentence Structure Checking: CheckForPlagiarism.net is said to have a sole advantage, that other soft-wares do not support, which is the fact that it uses a "patented" plagiarism checking approach. In which the sentence structure of a document is checked to ensure improper paragraphing and thus is susceptible to plagiarism. Also a synonym check is done to words and phrases to identify any attempt of plagiarism.
- Multiple Document Comparison: CheckForPlagiarism.net can compare a set of different documents simultaneously with other documents and can diagnose different type of plagiarisms at the sometimes [15].
- Supported Languages: CheckForPlagiarism.net supports English languages, Spanish, German, Portuguese, French, Italian, Arabic, Korean, and Chinese languages [15].

4.4 iThenticate

iThenticate one of the application or services designed especially for the researchers, authors' publisher and other. It provided by iParadigms that have introduced Turnitin in 1996 to become the online plagiarism detection. It is designed to be used by institutions rather than personal, but lastly they provided

a limit service for single plagiarism detection user like master and doctoral students and this allows them to check a single document of up to 25,000 words. So they can use this service to insure or to check their draft thesis whether containing correct citation and content originality [18]. The main features of iThenticate are [15]:

- Database Checking: iThenticate used its own database that contain millions of documents like (books, paper, essays, articles and assignments), with a large number of this documents that have been stored in iThenticate database locally, allowing the users who have account to do either online and offline comparison of submitted documents against it and to identify plagiarized content.
- Internet Checking : iThenticate, is considered as the first online plagiarism checker that provides live and cached links to websites and database to have extensive internet checking to all submitted documents. This Provides deep internet checking. One more advantage is that it can still check your documents even if a website is no longer online, this include all contents of website like forums, message boards, bulletin boards, blogs, and PDFs etc., all this check is done automatically and in (almost) real-time.
- Publications Checking: iThenticate offers an online and offline detailed and depth checking most types of publication like documents, including, books, articles, magazines, journals, newspapers, website and PDFs etc.
- Synonym & Sentence Structure Checking: Not supported by iThenticate.
- Multiple Document Comparison: iThenticate offers two types of document comparison document to document and multiple documents checking against database and also direct source comparison word to word also.
- Supported Languages: iThenticate supports more than 30 languages, it mean that it supports most of languages likes "English, Arabic, Chinese, Japanese, Thai, Korean, Catalan, Croatian, Czech, Danish, Dutch, Finnish, French, German, Hungarian, Italian, Norwegian, Polish, Portuguese, Romanian, Serbian, Slovak, Slovenian, Spanish, Swedish, Greek, Hebrew, Farsi, Russian, and Turkish." [18].

4.5 PlagiarismDetection.org

PlagiarismDetection.org: an online service provides high level of accuracy result in plagiarism detection. Mainly designed to help the teachers and student to maintain and to ensure or prevent and detect plagiarism against their academic documents. It provides quickly detect plagiarism with high level of accuracy [19]. The main features of PlagiarismDetection.org [15]:

- Database Checking: PlagiarismDetection.org used it own database that contains millions of documents like (books, paper, essays, articles and assignments).
- Internet Checking: PlagiarismDetection.org is an online plagiarism detector, so it is mainly based on the internet checking and is faster in plagiarism detection, it does not support offline detection.

- Publications Checking: PlagiarismDetection.org offers the students and teachers to check their publication against the published document and support most types of publication.
- Synonym & Sentence Structure Checking: PlagiarismDetection.org not supports Synonym & Sentence Structure Checking.
- mMultiple Document Comparison: PlagiarismDetection.org does not support multiple document comparison but it takes long time to return the result.
- Supported Languages: PlagiarismDetection.org supports English languages and all languages that using Latin characters.

Table1: Summarization of the comparison according to the software features: Key of the table figure1: The following expressions denote that: ***** Excellent, **** Very good, *** Good, ** Acceptable and * Poor.

Table 1. Comparsion of the software

Features	PlagAware	PlagScan	iThenticate	CheckFor Plagia-rism.net	Plagiarism detect-ing.org
Database Checking (on-line and offline)	*****	*****	*****	****	****
Internet Checking	*****	*****	*****	*****	*****
Publication Checking	*****	*****	*****	***	***
Multiple document com-parison	*****	*****	*****	****	****
Multiple languages sup-port	*****	*****	*****	*****	*****
Sentence structure and synonym checking	****	**	****	*****	**

5 Conclusions

The comparison of the software shown that still now their no software that can detect or to prove that the document has been plagiarize 100%, because each software and tool has advantages and limitation, according to the following features and performance, (Database checking whether locally or online, internet checking whether is online or offline, publications documents checking and supported types, capabilities of multiple document comparison, supported languages and synonym and sentence structure checking) we ranked them as follows, starting from the best, PlagAware, iThenticate, PlagScan, CheckForPlagiarism.net and lastly PlagiarismDetection.org. Academic enterprises can use one of the above software in their detecting of plagiarism but, due the limitation of most of the software and importance of plagiarism detection to the academic fields we suggest some rules that can be used to limit or to reduce student plagiarism teacher should educated student about plagiarism and its impact, copy right, citation and ownership.

In future work we may want to extend our comparison to larger and more varied set of real-life data and to extent our comparison to include more textual plagiarism software like Academic Plagiarism, The Plagiarism Checker, Urkund

and Docoloc, also to extended to include code source plagiarism detection software according to Supported languages, Extendibility, Presentation of results, Usability, Exclusion of template code, Exclusion of small files, Historical comparisons, Submission or file- based rating, Local or web-based and Open source

References

1. Alzahrani, S. Salim, N. Abraham, A. Understanding Plagiarism Linguistic Patterns, Textual Features and Detection Methods, preprint 2011.
2. Brin, S. and Davis, J. and Garcia-Molina, H. Copy Detection Mechanisms for Digital Documents. In: ACM International Conference on Management of Data (SIGMOD 1995), May 22-25, 1995
3. Baeza-Yates, R. & Ribeiro-Neto, B. Modern Information Retrieval. Addison-Wesley, 1999.
4. Bao Jun-Peng, Shen Jun-Yi, Liu Xiao-Dong, Song Qin-Bao, A Survey on Natural Language Text Copy Detection, Journal of Software, 2003, vol.14, No.10, pp. 1753-1760.
5. Encyclopedia Britannica, http://www.britannica.com/EBchecked/topic/462640/plagia (last access February 7, 2011)
6. Hage, J. Rademaker, P. Vugt, N. A comparison of plagiarism detection tools, Technical Report UU-CS-2010-015,June 2010, Department of Information and Computing Sciences Utrecht University, Utrecht, The Netherlands.
7. Hoad, T. C., Zobel, J. Methods for Identifying Versioned and Plagiarized Documents. JASIST 54(3): 203-215 (2003)
8. Jain, K. Murty, M. N., Flynn, P. J. Data clustering: a review. ACM Computing Surveys, 31(3):264–323, 1999.
9. http://www.cs.queensu.ca/queensu.ca/TechReports/Reports/2007-541.pdf. (last access February 7, 2011)
10. Jadalla, A. Elnagar, A. PDE4Java: Plagiarism Detection Engine for Java source code: a clustering approach. IJBIDM 3(2):121-135 (2008)
11. Kustanto, C. Liem, I. Automatic Source Code Plagiarism Detection. SNPD 2009:481-486.
12. Lesner, B. Brixtel, R. Bazin, C. Bagan, G. A novel framework to detect source code plagiarism: now, students have to work for real! SAC 2010:57-58.
13. http://www.plagaware.com/about_plagaware (last access February 7, 2011)
14. http://www.plagaware.com/about_plagaware/application (last access February 7, 2011)
15. http://plagiarism-checker-review.toptenreviews.com/index.html (last access February 7, 2011)
16. http://www.plagscan.com. (last access February 7, 2011)
17. http://www.checkforplagiarism.net (last access February 7, 2011)
18. http://www.ithenticate.com/index.html (last access February 7, 2011)
19. http://www.plagiarismdetection.org (last access February 7, 2011)
20. Zechner, M., Muhr, M., Kern, R., Michael, G. External and intrinsic plagiarism detection using vector space models. In: Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse. pp. 4755 (2009).

Anti-plagiarism Software: Usage, Effectiveness and Issues

Conference Paper · January 2015

DOI: 10.15308/Synthesis-2015-119-122

CITATIONS
3

READS
904

3 authors, including:

Violeta Tomašević
Singidunum University
35 PUBLICATIONS **121** CITATIONS
SEE PROFILE

Dejan Živković
Singidunum University
43 PUBLICATIONS **256** CITATIONS
SEE PROFILE

Synthesis
International Scientific Conference of IT and Business-Related Research

ANTI-PLAGIARISM SOFTWARE: USAGE, EFFECTIVENESS AND ISSUES

SOFTVER ZA ANTIPLAGIJARIZAM: UPOTREBA, DELOTVORNOST I MOGUĆI PROBLEMI

Marina Marjanović, Violeta Tomašević, Dejan Živković
Singidunum University, Danijelova 32, Belgrade, Serbia

Abstract:
In today's digital age, with never ending advances of information technology, plagiarism of textual documents is becoming one of leading issues that causes wide concern in higher education and science. The ubiquity of Internet has created a plethora of possibilities for unacknowledged copying and paraphrasing of other people's work. This has grave legal and moral repercussions for the society and seriously undermines its system of values. This paper discusses different types of document plagiarism and examines methods for their detection. It also presents software solutions that implement particular plagiarism detection techniques. Following the overview in the first part, the paper focuses on the analysis of difficulties arising in the plagiarism detection process and points out to open questions that need to be solved. Moreover, it offers some principal suggestions for possible improvements.

Key words:
plagiarism, plagiarism types, plagiarism detection methods, plagiarism detection issues, plagiarism detection software.

Apstrakt:
U današnje digitalno doba, koje odlikuju konstantne inovacije u oblasti informacionih tehnologija, pitanje plagijarizma u tekstualnim dokumentima postaje jedno od glavnih problema u domenu nauke i visokog obrazovanja. Sveprisutnost Interneta pruža mnoštvo mogućnosti za nelegalno kopiranje i parafraziranje radova drugih autora. To sa sobom nosi ozbiljne zakonske i moralne posledice za celokupno društvo i ozbiljno podriva njegov sistem vrednosti. U radu se razmatraju različite vrste plagijarizama u tekstualnim dokumentima kao i metode za njihovo otkrivanje. Takođe, pominju se softverska rešenja koja implementiraju određene tehnike za otkrivanje plagijarizama. Nakon kratkog pregleda u prvom odeljku, rad stavlja naglasak na analizu poteškoća koje mogu nastati u procesu otkrivanja plagijarizma i ukazuje na pitanja kojima bi se trebalo pozabaviti. Štaviše, autori nude načelne predloge i sugestije u cilju unapređenja softvera za otkrivanje plagijarizma.

Ključne reči:
plagijarizam, tipovi plagijarizma, metode za otkrivanje plagijarizma, problemi vezani za otkrivanje plagijarizma, softver za otkrivanje plagijarizma.

Acknowledgments:
The second and the third author wish to express their sincere gratitude to the Serbian Ministry of Education, Science and Technological Development for its support to paper presentation (Research Project III44006).

1. INTRODUCTION

In very broad terms, plagiarism can be defined as the act of uncritical use of other people's work (writings, thoughts, ideas, inventions, *etc.*) without acknowledging the source. While plagiarism can be traced back to almost the beginning of human civilization, the Internet has opened up numerous new possibilities for plagiarism, thus making it a very tempting endeavor.

The growing trend of plagiarism has become a grave global issue that seriously undermines the society's value system. This is why numerous countries have intensified their efforts to cope with the rise of plagiarism based on the combination of plagiarism prevention and detection. Plagiarism prevention refers to raising the society's awareness about the plague of plagiarism to a higher level, along with the implementation of a range of measures that include media campaigns and development of deterring strategies, honesty policies and sanctions. Plagiarism detection implies identification of unacceptable similarity between documents, usually by means of some sort of software systems.

In the battle against plagiarism, experience has shown that prevention is more effective than detection in the long run. This is the case mainly due to the fact that only plagiarism prevention measures can fully or to a great extent eliminate plagiarism, albeit consistently applied within a long period of time. On the other hand, plagiarism detection methods can only decrease the amount of plagiarism, even though they may achieve such positive results in the short term.

Not considering moral and ethical issues, it seems that there are two main reasons why confronting plagiarism is so difficult. Firstly, plagiarism itself eludes a clear universal definition because the borderline between the plagiarized and authentic work can be surprisingly blurred. Namely, except for obvious cases of the cut-and-paste kind, one can use many plagiarism techniques to disguise genuine scholarly work. Secondly, all plagiarism detection methods usually rely on software tools, while plagiarism is practiced by humans. Assuming that the plagiarist's goal is to go undetected with the "intellectual theft", the plagiarism issue can be considered an artificial intelligence problem of how well the computer can simulate the human thinking process.

The remainder of this paper is organized as follows: Section 2 reviews the existing types of plagiarism, while Section 3 elaborates on the methods and software tools used for plagiarism detection. The main goal of the paper is to analyze plagiarism issues and reasons for their emergence, point out open questions, and suggest possible improvements, shall be further discussed in Section 4. The final section includes a short summary with conclusions.

E-mail: mmarjanovic@singidunum.ac.rs
DOI: 10.15308/Synthesis-2015-119-122

2. PLAGIARISM TYPES

Plagiarism can be divided into about 15 types (Park, 2003; Park, 2004; Hiremath & Otari, 2014; Kashkur *et al.*, 2010; Turnitin white paper). They are listed below based on their detection difficulty.

1. Clone plagiarism – Taking someone else's work entirely.
2. Copy-and-paste plagiarism – Copying large parts of someone else's work.
3. Re-tweet plagiarism – Contains correctly quoted text, but relies too much on someone else's work.
4. Recycle or auto plagiarism – Publishing the same work many times.
5. Find-replace or word-switch plagiarism – Using synonyms for words in someone else's work.
6. Hybrid plagiarism – Combining judiciously quoted and unquoted parts of someone else's text.
7. Mashup plagiarism – Copying material from different sources.
8. Error plagiarism – Using incorrect citation.
9. Aggregator plagiarism – No originality in the work, although it contains references to the original work.
10. Style plagiarism – Paraphrasing to the extent that the original text is unrecognizable, but the structure of both documents is similar (essential schemes, main arguments, or examples coincide).
11. Translation plagiarism – Translating someone else's work into another language.
12. Idea plagiarism – The main idea of the work is not original, but it is masked by the plagiarist's knowledge.
13. Graphics plagiarism – Using a figure or a picture without permission.
14. Source-code plagiarism – Taking the source code in computer programming.
15. Ghostwrite plagiarism – Contracting another person or website to produce the work for someone.

3. PLAGIARISM DETECTION

The extraordinary popularity of the Internet has enabled easy access to useful and credible information for use by everyone. At the same time, the Internet has taken the plagiarism issue to a higher level by making it extremely easy for uncritical use of other people's work and even for finding numerous services on the Web that will do scholarly work for someone else. Thus, the continued growth of plagiarism cases has drawn the increased attention to the anti-plagiarism tools.

In today's digital marketplace, one can find many software products that offer defensive solutions against plagiarism based on various techniques. The following list summarizes the most commonly used methods for detecting different types of plagiarism (Hiremath & Otari, 2014).

Text-based plagiarism detection methods. Kashkur *et al.* (2010), provide a classification of techniques for plagiarism detection of textual documents. Furthermore, Meyer and Stein (2006) described a heuristic method for style plagiarism detection. The method is based on finding stylistic inconsistencies in the document being checked for plagiarism. However, this approach can give false positive results in case the document represents a joint work with multiple authors. Anzelmi *et al.*, elaborate on a detection algorithm that uses SCAM (Standard Copy Analysis Mechanism) formula for the so-called bag of words analysis (Anzelmi *et al.*, 2011). Moreover, Hoad and Zobel (2013) suggested that one can use the fingerprinting method to estimate the likelihood of similarity when two or more documents are compared.

Citation-based plagiarism detection methods. Hiremath and Otari (2014) described the method that uses citations and references for plagiarism detection. The method is based on an estimate of the degree of similarity in citations and the order of the documents being compared. This method can give good results for detecting idea plagiarism, but not for the hybrid and error types of plagiarism.

Shape-based plagiarism detection methods. Meyer and Stein (2006) derived a formula for detection of improper use of a figure without permission. The method is based on figure shape recognition, but it is very sensitive to even small changes in figure shape.

Source-code plagiarism detection methods. Kashkur *et al.*(2010), present many algorithms for the source-code plagiarism detection based on Kolmogorov complexity and fingerprinting method. Lukashenko *et al.* (2007), described various methods based on finding patterns of the same variable names in programs and identifying similarities in the syntax complexity of programs.

Translation plagiarism detection methods. Gipp describes a method for detection of the translation plagiarism based on the citation pattern analysis (Gipp, 2014).

Moreover, Urbina *et al.* (2010), provide an extensive list of commonly used software tools to detect plagiarism. Here, we shall reproduce the list by dividing it into two tables as given below: the first one is free software, while the other one is commercial plagiarism detection software. For each software tool in both tables, the second column specifies whether the tool is Web-based or a desktop application (web, desktop). The third column represents the corpus of documents that is searched over when plagiarism detection is performed (the Internet, database, files). The fourth column gives the acceptable file format of the document submitted for check (txt, pdf, img, ppt, and html). Finally, the fifth column shows the form of the report obtained as a result of plagiarism detection. The results can be sent as a percentage probability that the submitted document is plagiarized (%), as a website link to the report (link), or as a list of suspicious documents similar to the submitted document (list).

Software	App	Corpus	File format	Report
Approbo	web	internet	txt, pdf, doc	%, link, list
Image Stamper	web	internet	img	link
DocCop	web	internet, files	txt, pdf, doc	%, link, list
Plagiarism Checker	web	internet	txt	link
WCopyfind	desktop	files	txt, doc, html	%, link
Jplag	desktop	files	txt	%, list

Table 1. Free plagiarism detection software

DOI: 10.15308/Synthesis-2015-119-122

Software	App	Corpus	File format	Report
iThenticate	web	internet, database	txt, pdf, doc, html	%, link
Turnitin	web	internet, database	txt, pdf, doc, html	%, link
Plagiarism Detect	web, desktop	internet	txt, pdf, doc, html	link
Docoloc	web	internet	txt, pdf, doc, html	%, link, list
EVE2	desktop	internet, database	txt, doc	%, link
Scriptum	web	internet, database	txt, doc	%, link

Table 2. Commercial plagiarism detection software

4. PLAGIARISM ISSUES

The issue of plagiarism is inherently associated with the manner in which creative work is produced. There are dishonest authors who intentionally try to steal other persons' work. This, of course, represents a blatant case of plagiarism. However, the authors frequently create their own work by following and imitating others. There is a great risk that the work created may turn into a non-authentic piece, which is why the authors must be aware of the fact that good intentions are not an excuse. The only way the authors can be sure that their work is authentic is by doing it entirely on their own and by giving proper credit to other people's ideas.

The use of software systems for plagiarism detection has both positive and negative aspects. The advantages of software plagiarism detection come from the mere assets of the technology itself such as speed, reliability, easy reporting, *etc.* Negative effects result from misunderstanding of the role of software in plagiarism detection process. This can be manifested in the following ways:

- Software is accepted to definitely decide on whether some work is plagiarism or not, but it only serves the purpose to detect similarity of the document contents. This is exacerbated by the fact that software plagiarism detection tools often use catchy names to associate their purpose to something which is a way out of their league;
- Consequently, the authors often check their work for plagiarism and after obtaining a negative result, they deny that their work is plagiarism. This argument may not be in agreement with other people's judgment on (non-) authenticity of the work. The possible solution for such conflicting situations is to inform the authors in advance about the proper role of the software.

Software tools determine similarity of the document contents by using different methods and produce appropriate reports. The results of these reports need to be taken with great care and human judgment is necessary to decide whether something is plagiarism or not. However, in doing so, the question of the correctness of the obtained results must be clarified. Firstly, different software can give diverse results for the same type of document plagiarism, and thus, it is not clear how to interpret the results. Secondly, different software can analyze a document for different types of plagiarism, which raises the question of what is more relevant and how to reach a final decision.

Some authors have exploited the wide availability of cheap, even free, software tools for plagiarism detection by checking their work against plagiarism prior to making it public. Such bad practice can lead to an absurd situation in which the authors desperately try to revise a non-authentic work until it passes the plagiarism check, not paying much attention to the quality that usually deteriorates. Furthermore, for those who know how particular software works and which methods to implement, it is tempting to adapt their work by trying to circumvent software methods and pass the plagiarism check. These problems stem from the easy availability of software plagiarism detection tools to both authors and referees. If the plagiarism check was only in the hand of referees, the authors would be paying much more attention to their creative process and the quality of their work. The very idea of using software to check one's own work arouses suspicion that the work is not authentic and that it is nothing more than a lame attempt to soothe one's conscience. There is no need at all to use software to self-check an independent work for plagiarism, even though it is possible for software plagiarism detection tools to show similarities with other works.

Given the fact that human brain is much more complex than the computer, the role of human judgment in the process of plagiarism detection is indispensable. Namely, when detecting plagiarism, humans use semantic and statistical methods to apply them to all kinds of information. Human intuition, hunch and experience are also very important. Unfortunately, transferring these main features of human intelligence into software is far from possible today, and it may never be. On the other hand, the main advantage of computers is reflected in the computers' ability to access and process a huge amount of data with astonishing speed (in the plagiarism detection case, the data is a corpus of documents for comparison). This intrinsic characteristic of computers should certainly be of great assistance to humans in detecting whether something is plagiarism or not.

The complex questions regarding the plagiarism issue might be better answered by introducing more order in the process of plagiarism detection. One possibility may be the establishment of certified organizations with exclusive authority to check for plagiarism. The organizations would obtain certificates based on their activity, which would entail a certain level of responsibility. The certified entities could include publishing houses, universities, schools, and agencies providing plagiarism detection services. In the process of plagiarism detection, certified organizations should use software tools that are not freely available. The results of preliminary plagiarism detection process would be then made available to a committee which would be responsible for making the final decision on whether something is plagiarism or not. This would be an effective plagiarism deterrent and would boost the authors' morale and their self-confidence, while reducing the damage caused by plagiarism.

5. SUMMARY

Even though the incidents of plagiarism can be found since ancient times, plagiarism has never been as widespread as today. The rapid development of the Internet has significantly contributed to the proliferation of plagiarism cases. In fact, new digital technologies have triggered more opportunities for uncritical use of other people's work, thus making such new forms of plagiarism harder to detect and control. This unethical practice has become so serious that its erosive and corruptive effects are felt in all spheres of society. That is why the efforts against plagiarism have been intensified through implementation of a range of measures that usually involve software systems. Thus, one can ironically note that the information technology represents both the cause and the solution to the plagiarism problem.

DOI: 10.15308/Synthesis-2015-119-122

The lack of a clear and universal definition of the concept of plagiarism makes it more difficult to effectively prevent the old problem. As described in Section 2, it is possible to list 15 types of plagiarism, but this number is constantly increasing with the advent of new technologies. In order to detect new types of plagiarism, methods for determining the similarity of the document contents are also adapted (Section 3). A growing number of anti-plagiarism software tools are also available today, and some of the most popular ones are compared in Section 3. However, it appears that the anti-plagiarism software per se cannot solve the plagiarism problem. Moreover, the software tools are often misused and their results are misinterpreted, and thus, new problems emerge as a result of uncontrolled and incorrect use of the software. Some of these issues shall be discussed in Section 4.

In this paper, we have argued that the existing practice aimed to cure the plagiarism problem relying mostly on software tools is questionable. It should be thus revised by keeping its positive elements (for example, that humans make a final decision on plagiarism or that emphasis is placed on preventive measures to raise social awareness), as well as by getting rid of those elements that may cause new problems (for example, that anyone can verify the document contents). Greater level of discipline, stricter deterrent rules and more responsibility in the process of solving the plagiarism issue would probably give better results in the future.

REFERENCES

Anzelmi, D., Carlone, D., Fabio, R., Thomsen, R., & Hussain, D.M.A. (2011). Plagiarism Detection Based on SCAM Algorithm. Proceedings of the International MultiConference on Engineers and Computer Scientists, pp. 272-277.

Asim, M., El Tahir, A., Hussam, M.D.A., & Snasel, V. (2015). Overview and Comparison of Plagiarism Detection Tools. Department of Computer Science, VSB-Technical University of Ostrava, 17., Ostrava - Poruba, Czech Republic.

Gipp, B. (2014). Citation-based Plagiarism Detection – Detecting Disguised and Cross-language Plagiarism using Citation Pattern Analysis. Berlin: Springer Vieweg Research.

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

Kashkur, M., Parshutin, S., Borisov, A. (2010). Research into Plagiarism Cases and Plagiarism Detection Methods. *Scientific Journal of Riga Technical University Computer Science, Information T and Management Science*. 44, 139-143.

Lukashenko, R., Graudina, V., & Grundspenkis, J. (2007). Computer-Based Plagiarism Detection Methods and Tools: An Overview. International Conference on Computer Systems and Technologies - CompSysTech'07.

Meyer, S., & Stein, B. (2006). Intrinsic Plagiarism Detection. 28th European Conference on IR Research, ECIR 2006 London, pp. 565-569, Springer.

Park, C. (2003). In other (people's) words: Plagiarism by university students – literature and lessons learned. Assessment & Evaluation in Higher Education, 28, 471-488.

Park, C. (2004). Rebels without a clause: Towards an institutional framework for dealing with plagiarism by students. *Journal of Further and Higher Education*, 28, 291-306.

Senosy, A., Fadhil, N., Maidorawa, A., & Salim, N. (2014). Shape-Based Plagiarism Detection for Flowchart Figures in Texts. *International Journal of Computer Science & Information Technology (IJCSIT)*. 6(1). DOI: 10.5121/ijcsit.2014.6108

Shiremath, S.A., & Otari, M.S. (2014). Plagiarism Detection-Different Methods and Their Analysis: Review. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*. 1(7), 41-47.

Si, H., Leong, V., & Rynson, W.H. (1997). CHECK: A Document Plagiarism Detection System. In ACM symposium on Applied computing SAC'97, pp. 70-77, DOI: 10.1145/331697.335176.

Urbina, S., Ozollo, R., Gallardo, J.M., & Aina, C.M. (2010). Analisis de Herramientas para la Deteccion de Ciberplagio. XIII International conference EDUTEC 2010.

The Plagiarism Spectrum: Instructor Insights into the 10 Types of Plagiarism. Retrieved 15.02.2015 from https://www2.nau.edu/d-elearn/support/tutorials/academicintegrity/pdf/Turnitin_WhitePaper_PlagiarismSpectrum.pdf

DOI: 10.15308/Synthesis-2015-119-122

International Journal of Library & Information Science (IJLIS)
Volume 7, Issue 1, January–February 2018, pp. 1–6, Article ID: IJLIS_07_01_001
Available online at
http://www.iaeme.com/ijlis/issues.asp?JType=IJLIS&VType=7&IType=1
Journal Impact Factor (2016): 8.2651 (Calculated by GISI) www.jifactor.com
ISSN Print: 2277-3533 and ISSN Online: 2277-3584

PLAGIARISM DETECTION AND PREVENTION: A STUDY

Ganesh Kumar Soni
Student, Department of LIS, University of Rajasthan, Jaipur, India

ABSTRACT

In this paper we discuss about Plagiarism. What is Plagiarism? Their types, which citation it will be avoid, this papers main features are how to detect Plagiarism and how to prevent it and also difference between Plagiarism and Copyright infringement, Plagiarism and Ghostwriting. This article saying about effective Plagiarism detecting software's and this basis identify that how much work is real and other is fraud.

Key words: Plagiarism, Plagiarism Detection Software, Plagiarism Checking, Copyright Infringement, Ghostwriting.

Cite this Article: Ganesh Kumar Soni, Plagiarism Detection and Prevention: A Study. *International Journal of Library & Information Science*, 7(1), 2018, pp. 1-6.
http://www.iaeme.com/ijlis/issues.asp?JType=IJLIS&VType=7&IType=1

1. INTRODUCTION

In the present age of information technology any information is not accessing so difficult. The user can access the information 24/7 days from anywhere in the world. But with the use of information, it's misuse is also increasing some authors are publishing work from other writers by their name with little changes, it called Plagiarism. Plagiarism involves submitting work created by a professional service (in whole /part) and used without attribution.

The word Plagiarism comes from the Latin word 'Plagiarius' which means "abducting or kidnapping", it literary means theft, taking material authored by author's and presenting as someone else. Plagiarism is the illegal and unethical copying of another's work, which is up as its own.

In other words, acts of innocous quoting and borrowing become criminal or at least unethical, when the debt of one author to another is not properly paid via credit evidently due the receiver.

Unfortunately, digitization made copy-peste Plagiarism and inappropriate reuse of source from the websites, online journals and other electronic media. Widespread "with in academia, Plagiarism by students, professors and research scholars is considered academic censure, up to and including expulsion" and researchers and professors usually where punished for Plagiarism by sanctions ranging from suspension to termination with losing their credibility and perceived integrity.

2. TYPES OF PLAGIARISM

1. **Delibarate Plagiarism** - Deliberate Plagiarism is the simple and totaly wrong act of attempting to pass off someone else's work as your own.
2. **Paraphrasing** -This type of Plagiarism is a little more strategic. It involves reading a few texts, writing down a few key sentences, changing tha words around, and throwing in a few quotes and citations to throw your turor off the scent. Then you have the perfact essay, Right? Wrong!
3. **Patchwork Paraphrasing** - This is much the same as the above, except thet it involves reading from more books and 'Pathching' their ideas together.
4. **Bluffing** - This type of Plagiarism is Bluffing in the worst way, because you are pretending that you have ownership of certain ideas in order to fool others into thinking you know more than you do. In this Plagiarism reading books, journal articles, reports and reveals them in a new idea and is shown that they are different from them, but reality these thoughts are the same.
5. **Stitching Sources** - Stitching Source Plagiarism is within the "grey zone" because all the sources used are generally correctly cited, but the student has unsucceed to grow up their analytical power to enable them to work effectively and produce work that is truly their own. This is still Plagiarism, but is more likely to be accidental as a result of experience.
6. **Self-Plagiarism** - Self-Plagiarism refers to reuse his own work. The author can piece of his old work together with his new work and produce a successful amalgam. But can't get two grades for the same things/works, so even if you do this without knowing that it's wrong. It is still classed as Plagiarism.

3. PLAGIARISM V/S GHOSTWRITING

Plagiarism and Ghostwriting are related kinds of conduct in which the true author’s name is concealed the distinction between Plagiarism and Ghostwriting.

A Plagiarism copies text without permission of the true author.

A Ghostwriter knowingly and willingly produce text to appear as someone else’s speech or writing.

A Plagiarist does not pay the real author for his services, while a ghostwriter is always give a payment for his work.

The academic community make no distinction between Plagiarism and Ghostwriting. Either way, the name of the true writers is deceitfully covered by the students and student represent to have written text that he did not write.

Ghostwriting is not so good, because all over text in article is from unidentified author, while routine fraud involves, some original writing selection of sources by the students who submitted the work.

Plagiarism superior from Copyright infringements

Many legal judgment have equated Plagiarism with copyright infringement but this concept is incorrect when we see the following three aspects --

- In copyright law the doctrine of fair use allows an author to copy small amount of text (small sentences or whole paragraph) without the need for permission from the copyright author. In Plagiarism, if any word/text used in the article have been indicate in a quotation mark. When some words are copied without declaration of a quatations. The wrong in not copyright infringement, the false is failure to credit the words to the real author. However fair use will not secure the Plagiarists who copies much pages from a work into the Plagiarist's suspicious work.

- In the copyright law one cannot break text that is in the public domain (Like copyright has expired, author disclaimed copyright, work of Govt) but in Plagiarism it's always prohibited to copy material and non-copyright text is also prohibited without the conformation of quotation.
- Copyright law protects no one facts and none of ideas in the copyright work. Copyright protect only for expression of idea but some Professional Societies, Universities, Center's include copying idea in their explanation of Plagiarism

According to above quotations we realise that Plagiarism is not copyright infragment always and every condition. The embracement of real authors name and bibliographic data the copying not Plagiarization, but copying is copyright infringement.

How to avoid Plagiarism

Many people know that Plagiarism is unethical, unfair and crooked activity and this is usually enough to prevent us from doing it. Many technics are available to avoid Plagiarism but for those who do anything, the outcomes can be unpleasant, many professionals lost his professional reputation when they Plagiarize.

There are discussed some simple steps while writing research papers to ensure that your document will be free of Plagiarism--

- **Paraphrase:** When you find any information that is correct for your research article then first read it and add it into your own language. Make sure that you don't copy literally more than two words in a raw from the found text. If you use more than two words in a row then you will use quotation marks.
- **Cite:** Citing is a most dynamic mode to avoid Plagiarism. Going with the document formating instructions like APA, MLA Chicago etc. used by your educational institution. This generally require the addition of the authors and the date of publication or corrective information. Not citing properly can make Plagiarism.
- **Quoting:** Use the reference correctly the way it appears. No one wants to be misquote. Most educational communities of higher learning frown on "Block quotes" or reference's of 40 words or more. A student should be able to dramatically paraphrase most material. The reference's necessary done correctly to avoid Plagiarism allegations.
- **Citing Quotes:** This practice generally denote addition of page number, paragraph number in citation of web content.
- **Citing your own material (Self Plagiarism):** If your research article represent you own words/ideas, which related from your ongoing project, an earlier one or anywhere else you must cite yourself. Use the content as you like it someone else wrote it. It may sound odd but using information you have used before is called self-plagiarism and it is not delightful.
- **Referencing:** it need to use reference page of works at the end at your research article and this page in some as the document formatting instructions used by your research institutions. The reference are include the author details, date of publication, title and source. Follow the direction and get the right reference.

Why Plagiarism detecting

Plagiarism has a serious problem in present era. There are many examples of double published articles (in different journals) papers, and thesis. Plagiarism not only unethical, unfair, but also creates a problem for the real author, once he situate his work in another articles.

Unfortunately, when one reports a case of censure, one can't be sure of its importance. A search of the Internet indicates that there is more concern about cheating among students than among faculty. When title were scan, 90% of the article dealt with students and detection method of Plagiarism. This lack of articles on "Plagiarism among college faculty" could

indicate either that there is little Plagiarism among faculty or they are not willing to admit that there is a program of Plagiarism among faculty.

The most way to detect Plagiarism is to use online available tools, mainly software and online Plagiarism Checking Services. We are going to describe some such Plagiarism Detecting and Preventing softwares which are commonly use -

1. **Safe Assign**: Safe Assign provided by 'Mydropbox'. Safe Assign is based on a unique text matching algorithms capable of detecting exact and inexact matching between a submitted papers and source material, students submission to safe Assign are compared against several sources --

- Institutional Document Archives.
- Global Reference Database
- Pro Quest Journal Database
- Internet

Safe Assign a system that can perform both local and Internet detecting.

2. **Docol©c**- This Internet service provides by Institut fur Argewandte Lentechnolgien (IFALT). This plagiarism finding tools is searching for text fragments also available in other documents. Documents will be uploaded to Docol©c for an extensive reviews by software programme and large database. As a result you get plagiarism, copyright infringements, quotations or other sources of the documents in the web.

There are three step's to use it - (1) Log in (2) Upload Paper (3) Download Report

Docol©c provides the open access plagiarism search (OAPS) project. This type online service not any more available.

3. **Plagiarism Finder**:- It installs in to the user's computer and searches the internet for possible occurrences of text fragment from the local document collection. It detecting duplicate contents not only online tools, but also to it on proprietary database.
4. **Duplichecker**:- Duplichecker is the free online plagiarism checking software developers of this software tells like that it is 100% accurately Duplichecker software provide a statistical result of each text scan by user and provide a comprehensive analysis of his text.
5. **Viper Plagiarism Checker:** - Viper is a web based piece of software. Viper is easy to use interface and highly detailed scanning process. It only takes three simple steps- Check, Scan and Compare, with 10 Billion sources (including books, journal articles, websites and much more) to review your document and produce a full Plagiarism report. It launch in 2007, with a sample 'Drug and Drop'. Screening interface, easy to understand scan result and a free option available.
6. **Plagiarism Detector**:- Plagiarism detector is a software to detect plagiarism in online and offline documents this software mostly used to check for plagiarism in text document. It is fast a reliable plagiarism checker! This software detect the 1000 words online at a time. It's mainly task is the automatic detection of digital plagiarism (That is unofficial copy-paste or textual facts) that originated from the world wild web.
7. **PlagTracker.com:** - PlagTracker.com is the perfect free web based tool. PlagTracker is completely free to use. PlagTracker is a multifaceted online tool that meets the diverse needs of students. Turnitin and PlagTracker has the same operations but PlagTracker is available free of cost.
8. **WriteCheck:** - WriteCheck is online plagiarism tool its provide plagiarism checking by Turnitin ®, grammar checking by ETS ® technology and profession tutoring by personal tutor services. It also provide resource center for checking your essay/paper and point in the right direction, plagiarism quiz or understand about plagiarism, services.

Write check mainly worked for student to check grammar, style, usage, mechanics, spelling and originality.

9. **Glatt Plagiarism services:** - This software designed by Dr. Barbara Glatt. In this service Three program added for expose and shutout plagiarism.

- Glatt Plagiarism Teaching Program (GPTP) for define direct and indirect plagiarism and how to skip it.
- Glatt Plagiarism Screening Program (GPSP) for detect plagiarism and explain that how to different from copyright infringement.
- Last, Glatt Self Plagiarism Program (GSPP) a shield program to help detect unthought show of plagiarism.

10. **Plagium:** - Plagium is a free online multilangual plagiarism tracker. Its only 'Copy and Paste' online scanner, no one file type like, .docx, .PPT, .pdf, supported. Its provide quick search service for plagiarism, deep search for depth plagiarism checking and file comparison, for compare with uploaded files and URLs. Weekly alert service for regular users are also available. This software is able to handle check with large blocks of text. This software allows 2000 words per search at a time.
11. **iThenticate:**- iThenticate is developed by Turnitin. iThenticate is the leading provider of professional plagiarism detection and prevention technology used world-wide by scholarly. Publishers and research Institutions that ensures the originality of written work prior to publications.

iThenticate helps editors, writers, professionals, scholars, prevent misconduct by comparing manuscript against its database of our 60 billion web pages, 155 million content items including 49 million works from 800 scholarly publisher participates of Cross Check, a service presented by CrossRef and powered by iThenticate Software.

12. **URKUND:** - URKUND is a fully automatically system for handling plagiarism. When a student submit his thesis/dissertation. The Supervisor use URKUND Software to detect plagiarism

URKUND matches submitted data with three different sources:-

- The Internet
- Previously Submitted Student's Data (33 + Million, October 2017)
- Published material (Books, Journal, Report's etc.)

If URKUND find any similarity with above three content, it will flag it for possibility of plagiarism. Then the system sent a mail to the supervisor with his analysis.

13. **JPlag:** JPlag is a web service that finds pair at similarity program among a given set of programs. JPlag has a powerful user interface for understanding the results. JPlag is resource efficient and scales to large submission. Java is easy to use but it also available in C, C++, and Scheme.

Strategy for Detection

- As you read the articles, look for enclose evidence that may be point to plagiarism. Among the clues are following -

A) Mixed citation styles

B) Absence of quotations or references.

C) Paper Formatting

D) Outside Topic

E) Oddity with Style or Spelling

F) Smoking Guns.

- The Source of papers/articles must be read.

- Search URLs if paper online.
- Use plagiarism detectors like software, Plagiarism Resource Centre etc.

4. CONCLUSION

After studying the measures it is said that plagiarism is a crime. It will be happen then do not describe the author name and his literate details and if information will be online then must have URL.

Plagiarism mostly seen in the students. Students use the ideas, words of any author in their dissertation, projects, thesis without any proper citation. Students should be made aware of this for plagiarism prevention.

There are some chances that some users may use copy-paste method from the site. Plagiarism detection give permission to keep your essay antiplagiarism and checkout if it is also available somewhere else on the web are not. To check the reality of the content of your work such as an article, poem or essay, use a high quality Plagiarism Checker to find out if your content is Plagiarise or not. This is the citation that is common for library professionals, research scholars and students also.

Firstly, organization create a strict policy on plagiarism such as that of ACM (Association for Computing Machinery) this policy define about plagiarism, self-plagiarism and as well as define punishment. Legal action public humiliation and fines include in punishment criteria.

We must understand that failing this serious obligation will have dire consequences for the future social and the economic wellbeing of the world. Therefore, Plagiarism is a problem that must not be overlooked or swept under the rug.

REFERENCE

[1] http://www.scanmyessay.com

[2] http://en.writecheck.com

[3] Stanler, Ronald B.(2012).Plagiarism in colleges in USA.(www.rbs2.com/plag.pdf)

[4] http://www.plagiarism-detector.com

[5] http://www.duplichecker.com

[6] http://www.turnitinUK.com/

[7] http://www.plagiarism.com/

[8] http://www.iThenticate.com/

[9] http://www.URKUND.com

[10] http://help.blackboard.com/

[11] http://kmbigalk.tripod.com/

[12] Prechelt, Lutz. (2000). JPlag: Finding Plagiarism among a set of program. G Germany.p.5

[13] http://www.Wame.org/

[14] Kljajic B. Rjecnik Stranih rijeci.Zegreb:Nakladni Zavod;1990.p.1052

[15] Marsh, Bill (2007).Plagiarism: Alchemy and Remedy in Higher education. State university of New York.

[16] http://www.plagium.com

[17] http://www.jplag.ipad.kit.edu

[18] http://www.plagiarism-checker.com