Journal of Information & Knowledge Management
Vol. 16, No. 2 (2017) 1750011 (21 pages)

DOI: 10.1142/S0219649217500113

World Scientific
www.worldscientific.com

An Analysis of the Impact of Introducing the Plagiarism Detection System in an Institute of Higher Education

Ranjeet Kumar* and R. C. Tripathi†
Indian Institute of Information Technology Allahabad, Deoghat Jhalwa, India
*ranjeet@iiita.ac.in
†rctripathi@iiita.ac.in

Published 31 May 2017

Abstract. In the current technical scenario of the world, the importance of Intellectual Properties is growing at a fast pace in the business, education and the publishing industry particularly for those who have utilised the power of the internet to access any data, anytime, anywhere. The protection of copyright has therefore become a major concern in the said domain scenario since the theft of textual contents known as "Plagiarism" has emerged as a major menace. For this, a software plagiarism detection tool was developed earlier by us to find contents of a newly created work to be copyrighted in regard to what of its portions have been plagiarised and to what extent along with from which source whether from a local repository/database or from resources available over the internet. In the present paper, it has been used to analyse what way plagiarism would have infested the technical creativity of a typical newly established institute of higher education and research in IT, ECE, MBA, etc. had there been no such plagiarism detection tool available at all. The analysis is presented for arriving at overall scenario of plagiarism found in MTech thesis works, research papers, the book chapters written by faculty members and the research papers received in a couple of International Conferences organised by the institute etc. It is concluded that about 8.7% works manifested plagiarism of 20% or more in an institute just established almost a decade back. The tool has made profound impact in controlling plagiarism in the institute finally to reach almost zero level by fully revealing the contents plagiarised along with their sources. Guidelines have been arrived and reported here first time to advise the young authors how to mitigate the problem of plagiarism and thus assure that all the copyrightable works of the institute become free from plagiarism before they leave the station.

Keywords: Plagiarism detection; textual similarity; content similarity; intellectual thefts; copyright infringements/violations.

1. Introduction

With the advent of internet and the rise in use of computers among the common masses, plagiarism in various educational institutions and publishing industry is becoming a serious problem to deal with. A survey in June 2005 conducted as part of Centre of Academic Integrity Assessment project revealed that 40% of college level students admitted to have committed plagiarism as compared to 10% reported in 1999 (McCabe, 2005).

Plagiarism is the "wrongful appropriation" and purloining and publication of another author's "language, thoughts, ideas, or expressions", and the representation of them as one's own original work [Wikipedia Definition].

Plagiarism means copying someone else's work and claiming it as one's own work without giving due credit or by obtaining permission of the real owner or even excluding the other's work in the references section or by way of usurping any other credit which was due to the real owner viz. author/publisher. Today many projects, assignments, implementation reports whether of consultants, research organisations or the students of higher educational institutions, are found copied to various extents from the internet. With volume of material and text increasing on the internet annually by about 33%, most cases of copying may go unnoticed due to the absence of a robust plagiarism detection system. As a result, many copyright infringers may gain false credit for the work which is not their own.

The issue of plagiarism is now well known and very well defined in many papers. In this regard, after the offense is determined by some authority, the administrative actions for cases detected vary from case to case and from institution to institution. Table 1 gives an outline of typical Indian cases of plagiarism/unethical authorship reported in last decade along with punishments imposed on plagiarists.

On the line of Table 1, some recent plagiarism cases from other countries have been summarised in Table 2.

In the present academic scenario, the Institution of Electrical and Electronics Engineers (IEEE) policy on the issue of plagiarism is an important milestone for consideration. In this policy, there are specifications of percentage plagiarism and the academic penalty imposed on the author(s) whose research paper is found infected with that extent of plagiarism. Table 3 given below is the summary of IEEE copyright policy for research papers received by them for publication (IEEE Plagiarism Policy, 2015).

In several advanced countries including the United States, many universities have well-defined policies to classify and deal with academic misconduct. They provide all information and rules regarding the facts right at the time of enrolment process and via information brochures. Universities like Yale University, U.C. Berkeley University, MIT and some other European Universities like Oxford University, University of Cambridge also follow well-defined rules and regulations and have specific terms and conditions to deal with plagiarism cases and academic dishonesty.

Because of the unavailability of effective tools for checking content plagiarism, the practice of "cut"/"copy" and "paste" by many authors is going ahead unabated. Our institution (Odhyan *et al.*, 2013) developed earlier a tool known as "Prior Art Cop" and obtained a Singapore Patent in October 2013 which can filter out plagiarism in a query document by comparing it with existing relevant documents on the internet and also from documents already present in any soft repository. It also pinpoints the source from where the content has been copied and to what extent. Both the content of the query document as well as the content of the source wherefrom it has been copied are displayed side by side in a columnar form to enable

Table 1. Some typical plagiarism related cases and the resulting administrative actions as reported for India in last decade.

The unethical authorship case	The charge	The findings	The administrative actions
Controversy of Kumaon University	*Fully plagiarised* paper published with earlier published paper of another author	Committee suggested that he has personally *done no harm* and it was his students fault	He resigned from Vice-chancellorship immediately (*February* 2003)
Controversy of National Centre for Cell Sciences (NCCS)	Rehashed the same set of data which they had published earlier (*Misrepresented data*)	Committee advised the author to *take back* his paper	After Internal Investigation by its ethics committee, the Indian Academy of Sciences *banned* the author from participating in their activities for *three years (November 2010)*
Controversy of Sri Venkateswara University (SVU)	The author was accused of *Plagiarising* more than 70 published research papers	University executive Council *banned* him from undertaking examination work and research guidance	He was *barred* from securing further promotions and appointments to administrative positions (2004–2007)
Anna University Controversy	Published paper found *plagiarised* with another earlier published paper of another author	The journal reported that the paper "does not plagiarise the results presented but *copied most of it word for word*"	The Anna university *barred* one of the authors from guiding any more doctoral students (*2007*)
Controversy of Institute of Life Sciences, Bhubaneswar	Serious concern related to the *accuracy of the data* presented in many papers by him	Highly unethical practices such as serial *self-plagiarism, data manipulation and falsification of results*	*He was issued a notice* about five research papers by the journal of *Acta Biomaterialia* (*June 2013*)

Source: Wikipedia web-link http://en.wikipedia.org/wiki/Scientific_plagiarism_in_India.
Note: For details of the above cases please visit the above web-link from where it is summarised in a tabular form in this present paper.

anyone compare and verify what has been copied from where and to what extent. In this way, one can keep an effective tab on defaulters who copy work from other's copyrighted works beyond provisions of "FAIR USE" and claim the work to be of their own (Universal Law Publishing Allahabad India, 2013). The tool is of immense importance to retain one's copyright on a genuine work else the same may automatically land into public domain.

Plagiarism can be of many types and can have many levels (Pandey *et al.*, 2009). The most nascent form can be defined by a person who copies some content exactly as it is in the original source. This is called DITTO copying category and is the

Table 2. Some typical recent plagiarism/unethical authorship cases and the resulting Administrative actions as reported for outside India.

The unethical authorship case	The charge	The findings	The administrative actions
German's embattled Education and Science Minister, Annette Schavan	Using foreign text passages without proper citation in her 33 year old thesis	The university panel found the Christian Democratic Union (CDU) minister guilty	She resigned from the ministerial post of Education and Science (February, 2013)
[]Romanian Prime Minister, Victor Ponta	Half of the Doctoral Thesis in law is duplicated from other texts	Investigation of the charges revealed the source of plagiarism	The resignation of the country's minister because of misconduct (June, 2012)
Hungarian President, Pal Schmitt	The 16 pages of text from a German author and charts from a Bulgarian writer were lifted and inserted in his thesis.	The Semmelweis University panel found guilty and stripped him of his doctoral degree	He resigned from the president post after stripping of his doctoral degree (April 2012)
Jonah Lehrer, New Yorker	Exposed as a self-plagiarist and fabricator	The Wall Street Journal corrected the duplicated contents in two papers and unpublished two that contained self-plagiarism	Jonah Lehrer lost his job at The New Yorker (2012)

Source: Wikipedia web-link http://en.wikipedia.org/wiki/List_of_plagiarism_incidents.
Note: For details of the above cases, please visit the above web-link from where it is summarised in a tabular form in this present paper.

Table 3. Summary of IEEE Copyright policy specifying administrative actions for different level of plagiarisms.

Level of plagiarism	Definition of level	Possible corrective action
Level one	Pertains to the un-credited verbatim copying of a full paper, or the verbatim copying of a major portion (> 50%), or verbatim copying within more than one paper by the same author(s).	Notice of violation in IEEE Xplore
Level two	Pertains to the un-credited verbatim copying of large portion (between 20–50%) or verbatim copying within more than one paper by the same author(s).	Prohibition from publishing in IEEE or periodical
Level three	Pertains to the un-credited verbatim copying of individual elements (paragraph(s), sentence (s), illustration(s), etc.) resulting in a significant portion (up to 20%) within a paper	Rejection and return of papers in review and queues

Table 3. (*Continued*)

Level of plagiarism	Definition of level	Possible corrective action
Level four	Pertains to un-credited improper paraphrasing of pages or paragraphs	Referral to the IEEE Ethics and Member Conduct Committee
Level five	Pertains to the credited verbatim copying of a major portion of a paper without clear delineation (e.g., quotes or indents)	Repeat offenders subject to increased penalty

Source: IEEE Plagiarism Policy (2015).

easiest to detect. However, some people do the copying intelligently so that chances of getting caught are minimal. They do not just copy the content in its ditto form but make major modifications in it so that it may prove difficult to detect plagiarism inflicted by them. They achieve their motives in various ways by shuffling the paragraphs, mixing the statements, adding/removing statements; change the tense and form of words, and replace words with their synonyms or by way of using antonyms. This is called "Indirect Copying" or "paraphrasing". As a result, it becomes rather difficult for a fully automatic computerised system to find similarity between the modified document and an already existing almost same given document or its portions. Our earlier mentioned software tool "Prior Art Cop" uses an intensive algorithm to detect such plagiarism as well among documents which are modified to the extent popularly known as "paraphrasing".

2. Prior Art of the Work

Most of the plagiarism detection systems developed before said "Prior Art Cop" (Odhyan *et al.*, 2013) has been based on selecting fingerprints of *k-grams* (Karp and Rabin, 1987). A *k-gram* is a contiguous substring of length k. Document is divided into k-grams, where k is a parameter chosen by the user. Each k-gram is hashed and some subset of these hashes is selected to be the document's *fingerprints*. However the *k-gram* technique may not be an efficient approach. The biggest disadvantage of using *k-grams* is that it is based on string matching and not comparing the semantics of the text. It only matches two texts that have similar strings and not two texts that have similar meaning using different words. So, if *k-grams* are being used to develop a plagiarism detection system, it can only detect plagiarism in documents that are exactly copied (DITTO Copied) from some source or are just slightly modified. In the higher education (Ashworth *et al.*, 1997) studies on the students' perceptions of cheating and plagiarism behaviour in academic work and assignment were investigated. In their work Gullifer and Tyson (2010) also explored the university students perceptions of plagiarism on a focussed group of students. In their research Park (2003) reported the issues of plagiarism in the higher education and the students behaviour about plagiarism in lessons and assignments. Walker (2010) and

Youmans (2011) reported the issues of plagiarism and the role of plagiarism detection software in higher education to reduce the plagiarism cases in the academics.

In the emerging scenario of the work done for the plagiarism detection systems, shingling techniques, similarity measure calculations and document images have been reported as the most prominent technologies. Shingling techniques used in various research experiments such as COPS (Brin *et al.*, 1995), KOALA (Heintze, 1996), and DSC (Broder *et al.*, 1997), take a set of contiguous terms or shingles of documents and compare the number of matching shingles. The comparison of document subsets allows the algorithms to calculate a percentage of overlap between two documents. This type of approach relies on hash values for each document subsection and filters those hash values to reduce the number of comparisons the algorithm must perform. In this technique, the major issues remain as the efficiency of the system. Several optimisation techniques were proposed to reduce the number of comparisons made. In the Heintze, technique only portion of the shingles are retained whereas in Broder *et al.* (1997) technique only every 25th shingle is retained. However this method affects the accuracy. Since no semantic premise is used to reduce the volume of data, a random degree of "fuzziness" is introduced to the matching process. However, these result in relatively non-similar documents even being identified as potential duplicates.

Many other techniques and methodologies were developed for the documents to documents comparisons with high efficiency. In the research Buckley *et al.* (1999) and Sanderson (1997) have proposed a document clustering work. It was almost similar to Salton *et al.* (1975) work of document clustering, in that they used similarity computations to group potentially duplicate documents. In this methodology, all documents are compared, that means each and every document is compared with every other and a similarity weight is calculated. In another research paper (Kumar and Tripathi, 2013, 2014, 2015) an analysis of three major techniques of plagiarism detections was reviewed along with their representation of similarity calculations.

The extent of plagiarism is indeed a significant feature for consideration. In the paper published by Maurer *et al.* (2006) a thorough analysis of the plagiarism problem and possible solutions were discussed. They divided the solution strategies into three main categories. The most common method is based on document comparison in which a word for word matching is made of the query document with each target document in a selected corpus which could be the source of the copied material. A second category is an expansion of the document check but where the set of target documents is "everything" that is reachable on the internet and the query candidate to be checked for is a characteristic paragraph or sentence rather than the entire document. The third category mentioned by Maurer *et al.* (2006), is the use of stylometry, in which a language analysis algorithm compares the style of successive paragraphs and reports if a style change has occurred.

Plagiarism of text similarity in the documents submitted for the publications on the web for the academic purposes is most important. Numerous authors have assessed the problem in their own way. They have advised that plagiarism may be

endemic. These include Buckell (2002) and Culwin and Lancaster (2000, 2001a, 2001b). Other authors focussing on this type of plagiarism include Austin and Brown (1999), Lathrop and Foss (2000).

In the experiment of fingerprinting approach Manber (1994) focussing the n-to-n problem, a fixed resolution fingerprint is used first for the indexing of the collection, and then full fingerprinting is carried out for the query document. The said authors tested the results on collection of over 20,000 "read me" documents, identifying 3620 groups of identical files and 2810 groups of similar files. In line with the above work of fingerprinting, Brin *et al.* (1995) have developed a system called *COPS*. In the experiment they used hashed breakpoints strategy in the fingerprinting approach and also used a granularity of one sentence. The result was tested and the method extends the phrase if the hash value is equal to a multiple of a constant, k. The result was tested and it was found that correct documents scoring is arrived at on an average of 52.9%. In another experiment by Shivakumar and Garcia-Molina (1995, 1996) another system has been developed termed as *SCAM*, which used a granularity of one word, as opposed one sentence granularity by the *COPS*. As per author's statement, the results are improved over *COPS*. Later on they tested the improved version of the *SCAM* system on a large scale gigabyte database.

In another experiment of fingerprinting system Chowdhury *et al.* (2002) proposed an intelligent system called *I-Match*. It filters out terms based on inverse document frequency. In the pre-processing of the developed system, the most frequent and rarest terms are both removed. They used ranking method and weights based on the importance of document terms according to the frequency. The system focussed on detecting near exact copies. The system developed by Osman *et al.* (2012) based on semantic labelling and application of soft computing detects near exact copies.

Another well known system for plagiarism detection worth noting was based on feature vectors. Ottenstein (1976) came up with such a system in 1976. Later some more systems were reported but their performance was below the mark. The system developed by Schleimer *et al.* (2003) named MOSS (Measure of Software Similarity) used for detecting plagiarism in computer software codes used Rabin–Karp Algorithm with Windowing. However it was not of much use for detecting content plagiarism as it failed to detect the semantics associated with the document.

The system of our institute (Odhyan *et al.*, 2013) has the distinction that it can test plagiarism both for "ditto copying" as well as "para phrasing type" from existing documents on the internet as well as from the local repository of documents. On the other hand the majority of software programs currently available perform only one of the above two tasks.

3. Working Principle of Plagiarism Detection System of Our Institute

Our system (Odhyan *et al.*, 2013) was designed to include a strong architecture and a specific algorithm detect the different type of plagiarisms with their sources available

on internet or local repository. The copy and paste, paraphrasing, and semantic type of plagiarisms have all been addressed. As may be seen the techniques described earlier in the prior art section of this paper do not deal with all these types of plagiarisms in one go.

3.1. *The system workflow*

The system works in five steps as outlined in Fig. 1:

(1) First the user uploads the query document to be checked for plagiarism. The query document is stored in the local database.
(2) Intelligent and important queries are generated from the query document after preprocessing for stop word removal, stemming and hashing of both the query as well as repository documents.
(3) The queries generated above are searched on URLs of the internet and relevant pages of the search results are stored in the separate local database. Top Searches for each of the queries are saved and these act as the documents with which the system will compare for plagiarism in next run as these have the maximum chances of being in resemblance with the given query document.
(4) Using the URLs stored above, the system downloads the web pages from the internet with the hope that these documents may be similar in the context to the document being checked for plagiarism.

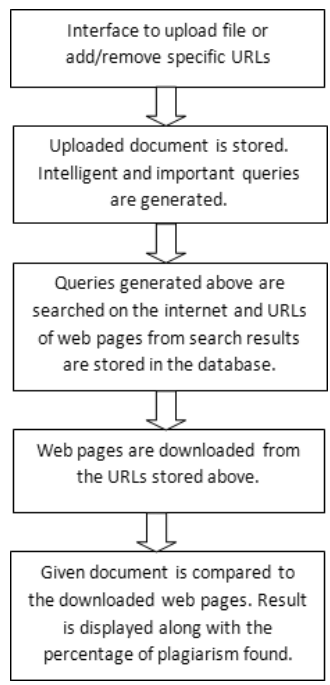


Fig. 1. An schematic work flow diagram of our prior art cop (Odhyan *et al.*, 2013).

(5) In the final step, the system compares the query document with all the documents downloaded from the internet or lying in the local repository. The result is now shown as a bi-columnar display of plagiarism detected in respect of the given document. The result shows the part of the document plagiarised along with the source URL from where it has been copied. This step is directly run after step (2) in case the search regime is limited to a local repository.

The entire system mentioned above is fully automated and the task of the user is limited to just supplying the query document which is required to be checked for plagiarism. Herein also, a most user friendly "Graphic User Interface (GUI)" is provided for the purpose.

3.2. *Results and their analysis*

The presently developed said system was automated to have a user friendly environment. An example output as results generated by the system is given below in Table 4 along with the percentage plagiarism found out by our "Prior Art Cop" software tool (Odhyan *et al.*, 2013). The output of the system is displayed in side by side two columnar displays. The system output gives a very clear view of the plagiarised portions and the source from where the contents have been copied. In the end of the output section, the total percentage of contents as copied from the different sources is clearly mentioned.

The efficiency and output results of the system have been tested on several input documents of Conference Proceeding, Journal papers, Master Dissertations, Doctoral Dissertations and for book chapters submitted by the faculty members. All the aforesaid research documents were compiled from 80 PhD students, 135 MTech students and 50 faculty members of the institute along with a couple of International Conferences organised annually by the institution over a period of six years (2008–2013) during which it was made compulsory for all to obtain a "clearance" from the Anti Plagiarism Cell which housed the plagiarism detection tool. No copyrightable work was allowed to leave the station until its plagiarism was tested and removed/cleared. The results of the system were found most satisfying and effective and obtained due acceptance by the aforesaid user communities. In each year, around 500 documents were evaluated for plagiarism checkup. The percentage plagiarism found out varied from case to case. Even 2–3 extreme cases having almost 100% plagiarism were also detected annually. Some selected extreme plagiarism cases from total 1000 cases found in the last 18 months are reported in Table 5. The test results are summarized in Table 6. It is observed that in total 8.7% cases manifested plagiarism beyond 20% i.e. of objectionable proportions. In them about 1.2% cases are extreme cases manifesting plagiarism to the extent of 50% or higher. Figure 2 given below shows how percentage share of plagiarism cases falls down for cases of increasing plagiarised portions in different slabs of extent of plagiarism.

The Fig. 2 given below shows the resulting graph of these extreme cases of plagiarism found in the last 18 months. Whereas 5% cases of plagiarism were found to

Table 4. The plagiarism results in side by side two columnar display along with the percentage plagiarism found in a query research paper.

Plagiarism Result

Given Document	Plagiarised with
Line No 55 Taking the exchange of water molecules into account, ankush identifies all water networks, gives detailed information about maximum residence time of the networks and the inter-network interactions. This is made possible by direct measurements of networks from each frame, hence is free from the artefacts of algorithms that take only water density into account or those algorithms that neglect the temporal (frame-by-frame) information. In absence of temporal information, it is difficult to identify if two water molecules that seem to interact actually are seen at the same time; thereby would be a potential source of misleading conclusions.	http://http://www.geocities.ws/dviip/src/ankush-1.0.6.pdf**Line No 4** Taking the exchange of water molecules into account, **ankush** identifies all water networks, gives detailed information about maximum residence time of the networks and the inter-network interactions. This is made possible by direct measurements of networks from each snapshot, hence is free from the artifacts of algorithms that take only water density into account or those algorithms that neglect the temporal/snapshot-based information. (In the absence of temporal/snapshot-based information, it is difficult to identify if two water molecules that seem to interact actually are seen at the same time; thereby would be a potential source of misleading conclusions.)
Line No 192 For each of the "hit" BLAST identifies, between the query sequence and database, BLAST uses a dynamic programming methodology to extend the hits in both the directions. This extension of score terminates when arriving at a score which falls more than little distance below the best score known among all other shorter extensions. The extension process can be lazy, if many hits are identified, but have the benefit of allowing the "gaps" in alignments. Based on the number of mismatches, length, and size and number of gaps in final alignment, BLAST program assigns an "expect" value to hit, showing how many much alike scoring hits BLAST can expect to discover with similar size inputs by accidental chance.	http://http://www.cs.washington.edu/education/courses/csep521/07wi/prj/hogg_russell.pdf**Line No 158** For each "hit" BLAST finds, between the query and database, BLAST uses a dynamic programming approach to extend the hit in both directions. The extension terminates when reaching a score that falls more than some distance below the best score among all shorter extensions. This extension process can be slow, especially if many hits are found, but has the advantage of allowing "gaps" in the alignments. Based on the length, number of mismatches, and number and size of gaps in the final alignment, BLAST assigns an "expect" value to the hit, expressing how many similar scoring hits BLAST would expect to find with similar sized inputs by random chance.
	Document has 31% of Plagiarism

Table 5. The relative share of plagiarism detected in the last 18 months.

S. No.	Percentage plagiarism range	Number of cases detected	Overall % share in total
01	20–30%	50	5%
02	30–40%	15	1.5%
03	40–50%	10	1%
04	50 and above	08	0.8%
05	100%	04	0.4%
		Total 87	Total 8.7%

Table 6. The results obtained from each of five online available plagiarism detection softwares in respect of 10 test cases having 20–30% plagiarism.

Some online plagiarism checking products	Percentage plagiarism 10–20%	Percentage plagiarism 20–30%	Percentage plagiarism 30–40%	Percentage plagiarism 40–50% and above	Percentage plagiarism free
VeriGuide	04 40%	02 20%	02 20%	—	02 20%
DocCop (File Check)	02 20%	—	—	—	08 80%
Plagiarism Detect	03 30%	—	04 40%	—	03 30%
Plagiarism Tracker	02 20%	—	—	—	08 80%
Turnitin	01 10%	—	04 40%	05 50%	—

have misappropriated contents to the extent of 20–30%, the cases found to have word for word copied from the source were 0.4% only.

The above results were obtained for a typical institution conducting courses of BTech (IT), BTech (ECE), MTech (IT), MBA (IT) and related PhD programmes. The Anti-Plagiarism Cell was setup so as to make all research publications free from plagiarism. In overall at anytime the institute had about 1400 BTech students, about 500 postgraduate students, 80 PhD scholars and 50 faculty members. The

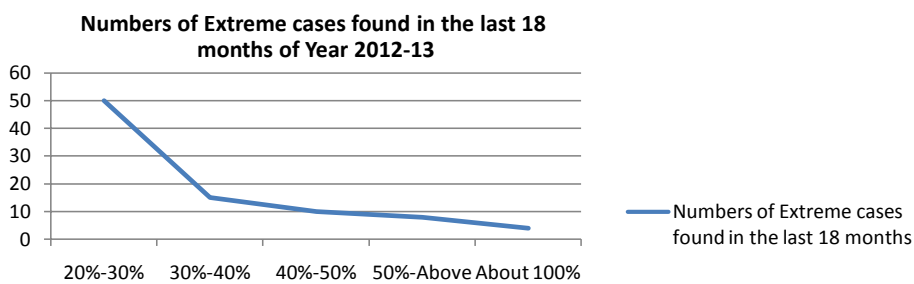


Fig. 2. The declining share of the extreme case plagiarisms detected in the query documents in the last 18 months.

overall trend of plagiarism cases found in the last six years (2008–2013) during plagiarism checkup process by this Anti-Plagiarism Cell has also been analysed. The data used in the current analysis is of a natural flow type with the objective to keep the research publications free from plagiarism so as to institutionalise and maintain the reputation of the institution. All the data used in the analysis was well documented and natural. It was free from any type of overrule or any biases.

Performance of our tool was compared with some other tools for same test cases. This comparison is given below in Table 6. In Table 6, out of 50 cases shown in Table 5 having plagiarism in the range of 20–30%, separate sets of 10 plagiarism cases (Research papers only) were tested. The above found as input query file for the well-known plagiarism detection tools is available online. They provide guest permission to check the plagiarism and their feature was used for the plagiarism checking process.

It is seen from the above table that the Veriguide, DocCop, Plagiarism detect and Plagiarism Tracker plagiarism detection softwares detected maximum number of such cases to have plagiarism in 10–20% or plagiarism free. Plagiarism Detect and Turnitin tools labelled their cases to have plagiarism mostly in the range 30–40% i.e. on the higher side from the real one of 20–30% as detected by our tool. Figure 3 shows the percentage plagiarism found from different software tools available online.

Table 7 shows the number-wise breakup of plagiarism cases for different extents in case of different types of copyrighted documents. Table 7 also shows the overall plagiarism checkup cases received and processed in over a period of six years along with their break up in terms of extent of plagiarisms found in them.

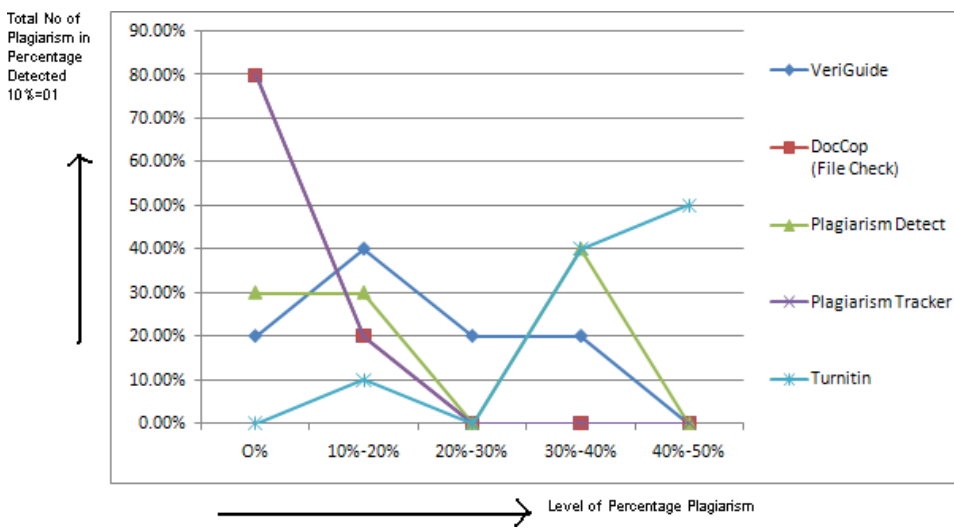


Fig. 3. The test cases results for our detected cases of 20–30% plagiarism as found out by five known plagiarism detection softwares online available.

Table 7. The estimated number of plagiarism checkup documents in last six years (2008–2013) in terms of maturity of the authors.

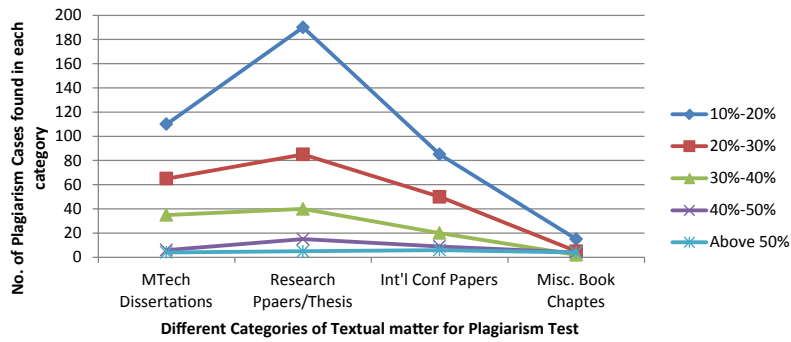
Document Type	Total No. of 6 Years	No. of Cases Free From Plagiarism	Break Up of the No. as per Plagiarism extent in % Ranges of plagiarism				
			10-20 %	20-30 %	30-40 %	40-50%	+ 50%
M.Tech Dissertations	650	450 [69.23%]	No % share 110 [16.92%]	No % share 65 [10%]	No % share 35 [5.38%]	No % share 6 [0.92%]	No % share 4 [0.62%]
Research Papers/Thesis	1100	800 [72.72%]	190 [17.27%]	85 [7.72%]	40 [3.63%]	15 [1.36%]	5 [0.45%]
Int'l .Conf. Papers	450	225 [50%]	85 [18.88%]	50 [11.11%]	20 [4.44%]	9 [2%]	6 [1.33%]
Misc. Books Chapters by Faculty	65	35 [53.84%]	15 [23.07%]	5 [7.69%]	2 [3.07%]	4 [6.15%]	4 [6.15%]
Total	**2265**	**1510 [66.66%]**	**400 [17.66%]**	**205 [9.05%]**	**97 [4.28%]**	**34 [1.50%]**	**19 [0.84%]**

The graph in Fig. 4(a) shows the relative number of plagiarism cases for different categories of copyrighted works in the institute over the said last six years. For every category of copyrighted work, the extent of percentage plagiarism in them declined constantly.

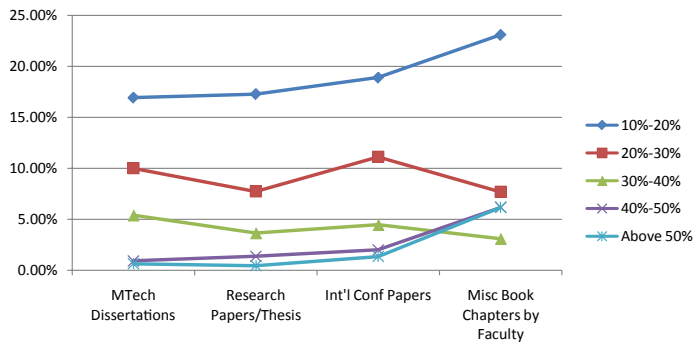
Accordingly Fig. 4(b) shows that the percentage plagiarism cases for different author categories in respect of different plagiarism extents.

The graphical representation of the composition of test data of the plagiarism checkup during last six years is given in Fig. 5. Figure 5 shows the typical annual test data breakup for the plagiarism checkup as received for various maturity levels of authors. In total of 2265 cases, 755 documents accounting to 33.34% (about one third) were detected to have significant percentage of plagiarism.

The statistical analysis of the plagiarism cases found in last six years in the institute is shown in Table 8, Figs. 6 and 7, respectively. They demonstrate the effect of the plagiarism checkup system put in place in a typical institute of higher education. It is observed that in the early phase when the system started working in late 2008, the plagiarism cases detected in the institute were very disappointing. However over the time, the percentage plagiarism cases dropped constantly. This drop is pronounced resulting in lesser extent of plagiarism indicating that the general authors took a note of being caught for plagiarism. However the share of those who opt for major chunk of "ditto copying" is not affected so much by the existence of in-house plagiarism detection tool. This speaks of their inability to respond to the



(a)



(b)

Fig. 4. (a) The number-wise plagiarism cases for different author categories in respect of different plagiarism extents. (b) The percentage plagiarism cases for different author categories in respect of different plagiarism extents.

developments in the environment on account of their hopelessness to adjust with the new developments.

For sake of comparison, a complete list of 450 documents received during the said six years period containing the research papers from various national and

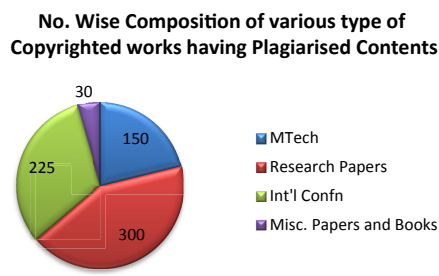


Fig. 5. Number wise composition of various types of copyrighted works from different levels of author maturities.

Table 8. The total number of plagiarism cases over the last six years.

Year	Total No. of cases	Percentage				
		10–20% plagiarism found	20–30% plagiarism found	30–40% plagiarism found	40–50% plagiarism found	50% and above plagiarism found
2008	230	95 41.30%	52 22.60%	28 12.17%	10 4.34%	05 2.17%
2009	325	85 26.15%	46 14.15%	22 6.77%	08 2.46%	03 0.92%
2010	355	75 21.13%	39 10.99%	18 5.07%	05 1.40%	03 0.84%
2011	370	60 16.22%	28 7.57%	13 3.51%	04 1.08%	02 0.54%
2012	390	46 11.79%	22 5.64%	09 2.31%	04 1.02%	03 0.76%
2013	595	39 6.55%	18 3.02%	07 1.18%	03 0.51%	03 0.51%
Grand total	2265	400 17.66%	205 9.05%	97 4.28%	34 1.50%	19 0.84%

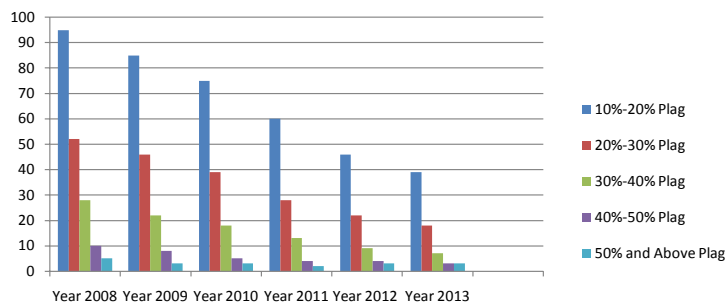


Fig. 6. The ratio of the plagiarism declining constantly over the years.

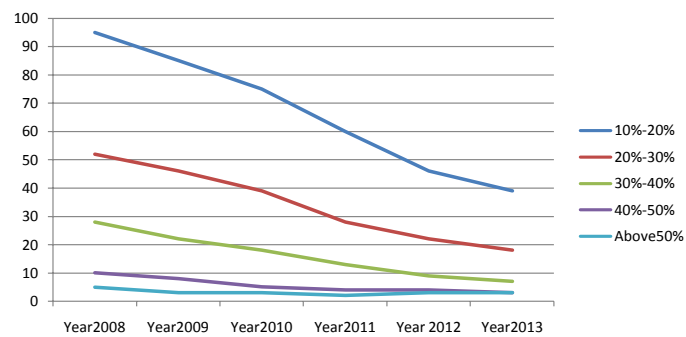


Fig. 7. The year-wise percentage plagiarism of last six years.

Table 9. The total number of plagiarism cases over the last six years from outside of the home institution.

Documents type	Total no. in six years	No. of cases free from plagiarism	Plagiarism 10–20%	Plagiarism 20–30%	Plagiarism 30–40%	Plagiarism 40–50%	Plagiarism above 50%
Research papers received from outside by the home institution	450	225 (50%)	85 (18.88%)	50 (11.11%)	20 (4.44%)	9 (2%)	6 (1.33%)
Home institution	1815	1285 (70.79%)	315 (11.84%)	155 (8.54%)	77 (4.24%)	25 (1.38%)	13(0.72%)

internationals academic and research institutions outside were also investigated. The data were from outside of the home institution and decision regarding their acceptance in the international conferences held at the home institution was handed over to the authors also for their concern and they accepted the facts found from our plagiarism reports. The data is shown in Table 9.

4. Discussion

A review of results arrived at in this paper reveals that in the decade of 1990s and thereafter the access to the internet and therefore easy copying of its contents allured many authors to opt for plagiarism resulting in increasing share of the plagiarism cases. The situation was also abetted because of lack of proper and efficient tools for detecting plagiarism and demonstrating it to a copyright infringer in respect of what he/she copied in unfair way in which part of his/her document, from where and to what extent. Over the years the plagiarism detection technology has grown from its early approaches based on keyword overlapping, shingling, visualisation and finally to the latest tool developed in our institute (Odhyan *et al.*, 2013) which not only detects the cases of ditto copying but also the cases involving various types of paraphrasing. Our tool is able to checkup plagiarism based on a windowing technique both on local document repository as well as from all the possible URLs of the internet accessible as such and concerned to the manuscript in question. It also gives a side by side bi-columnar tabular display of the final results showing the portions of a query document vis-à-vis portions along with the address of URL from where the contents have been copied and to what extent. This makes the system fully transparent displaying to the plagiarists to convince them as well as audiences what has been plagiarised leaving no room for the author for making any excuses in this regard. As a result various cases of plagiarism detected and discussed with the authors revealed their inherent ignorance of the law related to authorship as defined in the Copyright Act of the country and also his/her inhibitions related to concepts of drafting a research paper.

The present paper has been able to reveal some real time experiences gained as consequences of deploying of our latest software tool for plagiarism detection in a newly established institution vis-à-vis what could go unnoticed its number. plagiarism checkup tool was deployed. During the last six years of work experience, many issues were noted in regard to the plagiarism results. The research students were seen to advance typical reasons to defend the plagiarism inflicted by them. A deeper study of the same revealed lack of their knowledge how to avoid plagiarism in terms of the following most prominent types of these outlined as below.

4.1. *Handling of the well-known technical standards and facts*

Many amateur authors were found to have a rigid stand in regard to reproduction of well-known technical standards, natural laws and facts. Examples include the seven-layer networking architecture of OSI for data communications, sections of a country legal acts/laws, scientific principles and their formulae, etc. The authors pleaded that such items need to be reproduced ditto "as such". They had the feeling that such items being of universal nature cannot be expressed exactly by them in their own language. The authors therefore preferred to reproduce them in the body of their research paper as such.

To mitigate the above problems, they were advised to follow typical guidelines as available in the plagiarism policy of renowned professional bodies like IEEE. Accordingly, they were advised to have their summary expression as a portion in the body of their document and send the detailed version of such technical standards, laws and facts to the "Appendix" section of their research work. They were also advised that as per the copyright laws, if a typical para from a reference is necessarily to be reproduced within the body of their research work, this is allowed once they give the complete reference right at the place of reproduction and make the reproduced para indented and italicised.

4.2. *Self-copying*

Some of the cases found with heavy plagiarism by way of copy and paste involved same author's own publications with different colleagues/author(s) in different journals. It was found that the authors had a firm belief that any work which they have authored earlier is always under their own copyright and therefore can be freely copied by them in their future works. Accordingly they felt that if some research papers were published earlier by them with two or three different authors, they were authorised to reproduce the same for any journal where they want to submit their new work related to the same topic to be published. With great difficulty, they were finally convinced that any expression in the form of a research paper though of their own, once submitted to any journal is published only when the author has assigned his/her copyright to the publisher of the journal. Thus once they have filled in and submitted their earlier work to a journal publisher by way of "*Copyright Transfer Agreement (CTA)*", the same is no more their own intellectual property. This is in

line with the copyright laws of the country since same contents copyright which were once transferred to one publisher, cannot be given to another publisher for publication without prior and explicit consent of the first publisher. In case they want to reproduce the same in a book or any other journal, a prior permission of the first publisher whom they have submitted CTA is mandatory. Violation of such a practice amounts to copyright infringement for which civil and criminal remedies as specified in the Copyright Act of the concerned country may obtained be invoked through the judicial courts.

4.3. *Pretending for inadvertent copying*

Some cases of 100% plagiarism in the institute were also detected. For this the concerned faculty members implicated their students and vice versa. Both were noted for not confessing the truth. Rather, each one passed on the buck on the remaining coauthors. Some of them also came forward to the excuse that the CD containing the document [*Any document for plagiarism check is, first written on the CD and sent to the Anti Plagiarism Cell for Plagiarism Checkup Process*] sent for plagiarism checkup was swapped with another one containing copy of someone else's research paper copied for studies only. All these on due investigation were found to be fake excuses. A typical truth was found that in such cases, the concerned faculty member failed to assess the capability of his/her student(s). The student(s) also admitted this truth and under the pressure of course requirement that at least one research paper has to be published for final year MTech/MBA Dissertation, a few submitted a hundred percent copied research paper from some very remote and less current URL in the hope that the same will not be detected as a case of plagiarism. Remedy of such cases emerged not to impose mandatory publication of a research paper by all the final year Master Degree students. Rather, some alternate route needed to be provided for completion of the course work for obtaining the degree.

5. Conclusion

The present paper brings out the anatomy of the plagiarism cases and authors excuses their of in a newly established institute of higher education. Six years data (2008–2013) of plagiarism in a typical institute of higher education was analysed. Almost two third works were found totally free from plagiarism. In the remaining in overall 8.7% of works like for Master degree dissertations, PhD thesis works, research papers, papers of International conferences and book chapters written by faculty members, manifested plagiarism beyond 20%. About 1.2% cases amongst them were of extreme category manifesting plagiarism beyond 50% and 0.4% were of 100% copying. Extent of plagiarism in MTech thesis works which are examined in person by outstation experts and are joint efforts of a young researcher and a faculty is always less than that in individual research papers created primarily by research scholars. 15% to 20% of book chapters written by faculty are inflicted with

plagiarism to the extent of 10% to 20%. The share of plagiarism cases falls exponentially with extent of plagiarism in them. Deployment of an effective Prior Art Cop s/w tool of our institute along with counselling the researchers how to keep away from plagiarism is able to keep an institution free from evils of plagiarism.

The present study also revealed inherent inhibitions of the authors involving their ignorance and passing the buck type of excuses advanced when a real case of a plagiarism is detected and presented to the author(s). Experience and expertise of the authors to enable them understand their specific weaknesses and follow the laws of the copyright was revealed. Specific categories of most prevalent plagiarisms, like (i) reproductions of the standards, natural laws and facts; (ii) self plagiarism and (iii) pretending of inadvertent copying, were quantified first time in the scientific literature along with the remedies for them. The future scope of the present work seems to cover not only the textual plagiarism but also the plagiarism of tables, flow charts, block diagram and images.

References

Ashworth, P, P Bannister, P Thorne and Students on the Qualitative Research Methods Course Unit (1997). Guilty in whose eyes? University students' perceptions of cheating and plagiarism in academic work and assessment. *Studies in Higher Education*, 22(2), 187–203.

Austin, M and L Brown (1999). Internet plagiarism: Developing strategies to curb student academic dishonesty. *The Internet and Higher Education*, 2(1), 21–33.

Brin, S, J Davis, H Garcia-Molina (1995). Copy detection mechanisms for digital documents, In *Proceedings of the ACM SIGMOD Annual Conference*, San Jose, California, May 22–25, pp. 398–409.

Broder, A, S Glassman, S Manasse and G Zweig (1997). Syntactic clustering of the web. In *Proceedings of the Sixth International World Wide Web Conference (WWW6'97)* Santa Clara, CA, April, pp. 391–404.

Buckell, J (2002). Plagiarism tracked at 8 per cent, Australian IT, 9 November, p. 19.

Buckley C, C Cardie, S Mardis, M Mitra, D Pierce, K Wagstaff and J Walz (1999). The smart/empire TIPSTER IR system. In *Proceedings of TIPSTER Phase III*, (San Francisco, CA), pp. 107–121.

Culwin, F and T Lancaster (2000), A review of electronic services for plagiarism detection in student submissions. In *Proceedings of 1st LTSN-ICS Conference*, Edinburgh, pp. 54–61.

Culwin, F and T Lancaster (2001a). Plagiarism issues for higher education, Vine 123, available from LITC, South Bank University, London.

Culwin, F and T Lancaster (2001b). Plagiarism prevention, deterrence and detection. Available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.107.178&rep=rep1&type=pdf. Accessed on 23 April 2017.

Chowdhury, A, O Frieder, D Grossman and M McCabe (2002). Collection statistics for fast duplicate document detection. *ACM Transactions on Information Systems*, 20(2), pp. 171–191.

Gullifer, J and GA Tyson (2010). Exploring university students' perceptions of plagiarism: A focus group study. *Studies in Higher Education*, 35(4), 463–481.

Heintze, N (1996). Scalable document fingerprinting, In *Proceedings of the USENIX Workshop on Electronic Commerce*, November, Oakland, California.

IEEE Plagiarism Policy (2015). Available at http://www.ieee.org/publications_standards/publications/rights/plagiarism_FAQ.html. Accessed on 23 April 2017.

Karp, RM and MO Rabin (1987). Efficient randomized pattern-matching algorithms. *IBM Journal of Research and Development*, 31(2), 249–260.

Kumar, R and RC Tripathi (2013). An Analysis of automated detection techniques for textual similarity in research documents. *International Journal of Advanced Science and Technology*, 56(9), 99–110.

Kumar, R and RC Tripathi (2014). A trigram word selection methodology to detect textual similarity with comparative analysis of similar techniques, *4th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, 7–9 April, pp. 383–387.

Kumar, R and RC Tripathi (2015). Text mining and similarity search using extended tri-gram algorithm in the reference based local repository dataset. *International Journal of Procedia Computer Science*, 65, 911–919.

Lathrop, A and K Foss (2000). *Student Cheating and Plagiarism in the Internet Era — A Wake Up Call*. Englewood: Libraries Unlimited.

Manber, U (1994). Finding similar files in a large file system. In *1994 Winter USENIX Technical Conference*, San Francisco, CA, January, pp. 1–10.

Maurer, H, H Krottmaier and H Dreher (2006). Important aspects of digital libraries, *International Conference of Digital Libraries*, New Delhi, 5–8 Decmber, pp. 843–855.

McCabe, D (2005). The Center for Academic Integrity's Assessment Project Research Survey. Available at http://www.waunakee.k12.wi.us/hs/departments/lmtc/Assignments/McConnellScenarios/AcadHonesty_5Article.pdf. Accessed on 23 April 2017.

Odhyan, S, A Mazumdar, T Kumar, RC Tripathi and MD Tiwari (2013). A method and a software implemented tool for detecting plagiarism in documents. Hungry Patent Application Number 170694, Application No. 201007580-2 (2010).

Osman, AH, N Salim, MS Binwahlan, R Alteeb and A Abuobieda (2012). An improved plagiarism detection scheme based on semantic role labeling. *Applied Soft Computing*, 5(5), 1493–1502.

Ottenstein, (1976). An algorithmic approach to the detection and prevention of plagiarism. *SIGCSE Bulletin*, 8(4) 30–41.

Pandey, A, SS Math and RC Tripathi (2009). A case study of plagiarism detection tools and techniques. In *Proceedings of the National Seminar on e-Learning*, NASEL, Ewing Christian College, Allahabad, 21–22 March, pp. 1–4.

Park, C (2003). In other (peoples) words: Plagiarism by university students — literature and lessons. *Assessment & Evaluation in Higher Education*, 28(5), 471–488.

Salton, G, CS Yang and A Wong (1975). A vector-space model for information retrieval. *Communications of this ACM* 18(11), 613–620.

Sanderson, M (1997). Duplicate detection in the Reuters collection. *Technical Report* (TR-1997-5) *of the Department of Computing Science University of Glasgow*, Glasgow G12 8QQ, UK.

Schleimer, S, DS Wilkerson and A Aiken (2003). the winnowing: Local algorithms for document fingerprinting (MOSS). In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, San Diego, California, 9–12 June, pp. 76–85.

Shivakumar, N and H Garcia-Molina (1995). A copy detection mechanism for digital documents. In *Proceedings of the International Conference on Theory and Practice of Digital Libraries*, Austin, Texas, pp. 1–13.

Shivakumar, N and H Garcia-Molina (1996). Building a scalable and accurate copy detection mechanism. In *Proceedings of the ACM Conference on Digital Libraries*, Bethesda, Maryland, 20–23 March, pp. 160–168.

Universal Law Publishing Allahabad India (2013). The Copyright Act 1957 with amendment to date Ch XI sec 51 "Infringement of Copyrights".

Walker, J (2010). Measuring plagiarism: Researching what students do, not what they say they do. *Studies in Higher Education*, 35(1), 41–59.

Youmans, RJ (2011). Does the adoption of plagiarism-detection software in higher education reduce plagiarism? *Studies in Higher Education*, 36(7), 749–761.