An Analysis of Automated Detection Techniques for Textual Similarity in Research Documents

Ranjeet Kumar[1] and R. C.Tripathi[2]

Patent Referral Centre
[1,2] Indian Institute of Information Technology, Allahabad
[1]ranjeet@iiita.ac.in, [2]rctripathi@iiita.ac.in

Abstract

In the present age of Internet access, the world is growing with lots of easily accessible information available on almost every subject matter. The use of the freely available internet resource is causing easy copy and paste culture resulting in plagiarism in various research documents and academic reports. In such a scenario of the growing research and development publications, many techniques and methodologies have been developed for the plagiarism detection to evaluate the originality in the research documents both in regard to the web based as well as local repository based contents. Most important techniques and methodologies in use have been reviewed in this research paper, with their due analysis in terms of their methodologies, procedures and working principles along with their efficiency in finding the text similarity for a given query document.

Keywords: *Near exact Text matching, Text Similarity finding, Textual Plagiarism detection, Text Retrieval, Trigram Retrieval*

1. Introduction

Plagiarism is defined as borrowing some one else expression and presenting it as one's own. This is an offence under the "Copyright Act" of every country. In the academic, scientific and technical as well as industrial literature, there is ample chance of textual similarity of a query research paper/patent with other published or copyrighted texts available on the internet. This is so because in the current knowledge society, any type of researcher can find freely abundant relevant literature on the internet. The textual plagiarism is becoming a menace in the current internet age. Since almost everything seems to be available on the internet and an user can access the materials one gets tempted and can copy it easily.

The text similarity in the research documents causes plagiarism. The issue of the plagiarism in the academia and research organizations is today a prime focus. The textual similarity in the document plagiarism is considered to be a most serious scholaristic misconduct. Academia everywhere is undertaking efforts to educate the issues of the concerned plagiarism and how to avoid it. For the basic measuring criteria in the real practice to find whether two documents are similar or not, two basic approaches are now well established. These are i) the local (local repository or direct) and ii) global (in large set of data available on internet globally or indirect) approach (Ahlgren *et al.*, 2003; van Eck and Waltman 2009). In the local approach, the similarity between two documents is found by direct matching with the second documents in the local repository. In the global approach, the similarity between two objects is obtained by measuring the similarity between their profiles, vectors that often contain the number of co- occurrences (eventually normalized) of an object with each other considered as objects (Cristian Colliander, 2011). A number of papers in the scientometric literature report outcomes of comparisons of similarity measures or similarity approaches. Outcomes of empirical comparisons are reported by Boyack *et al.*, (2005, 2011), Boyack and Klavans (2010), Ahlgren and Colliander (2009a, b), Ahlgren and Jarneving (2008), Leydesdorff (2008), Klavans and Boyack (2006), Gmu¨r (2003), Luukkonen *et al.*, (1993), and Peters and Van Raan (1993). Other studies report outcomes of theoretical comparisons (Egghe 2009, 2010a, b; Egghe and Rousseau 2006; Hamers *et al.*, 1989). In reports by Van Eck and Waltman (2009), and in Egghe and Leydesdorff (2009), outcomes of both empirical and theoretical comparisons are reported.

2. The Prior Art Research

Detecting plagiarism and how to avoid academic plagiarism has been an area of interest since the long past. Denning (1995) suggested establishing libraries of academic works. The idea behind this is to avoid the plagiarism in the academic submissions in the Journals or Conferences for the publications of the research articles or books. This would seem to be a generally sensible idea, similar to the way in which detection services are of importance in areas like raw materials to the factories. Samuelson suggested that 30% shared similarity is acceptable for self plagiarism (Samuelson 1994). In the current situation however, this is not so good or clear cut. The similarity of the text should still avoid being the word by word or exact similarity. It could also be expected that similar background material would be there and similar to some extent in the beginning of many academic submissions. It is now not clear cut solution of the matter of 30 % that whether 30% means 30% of word count or 30% of page count. The pages would any way include diagrams, pictures, flow charts, tables *etc.* which are not counted as such in the textual plagiarism. Hence a more detailed examination of what constitutes self plagiarism is necessary to determine what appropriate prevention methods should deal with them.

The extent of plagiarism is indeed significant. In the article published by Maurer *et al.*, (2006) a through analysis of the plagiarism problem and possible solutions has been discussed. They divide the solution strategies into three main categories. The most common method is based on document comparison in which a word for word check 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 candidate to be checked for is a characteristic paragraph or sentence rather than the entire document. The third category mentioned by Maurer *et al.*, 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 or towards submission for the academic purposes is most important. Numerous authors have assessed the problem in their own way, have provided advice or stated that plagiarism may be endemic. These include Buckell (2002) and Culwin & Lancaster (2001a, 2001b). A particular concern of Web plagiarism has been identified. Authors focusing on this type of plagiarism include Austin & Brown (1999), Culwin & Lancaster (2000b) and Lathrop & Foss (2000).

It is worth noting that there are many websites and articles on the internet that mention about the plagiarism in general. The high level universities, research laboratories and Scientific Organizations all have certain rules, guidelines and mechanisms for finding the similarity in the written text for their publications. They use some software or some manual mechanism for the detection of the similar text which one can say “textual plagiarism” before the research articles or any books are published. There are various electronic software tools for plagiarism detection. In addition to whom there are many Web based plagiarism detection services which provide plagiarism report for the submitted query text document; of course for some payment.

One of the earliest plagiarism detection tools was propounded by Denhart. He recommended EVE over iParadigms (Denhart 1999). He suggested that EVE is the more suitable plagiarism finding tool and recommended that 'iParadigms' is very quick to say that a document contains plagiarism, when it should only say that some similarity has been found, since, some of the document submitted was legitimately fully cited in the literature. In the later stage and more advanced research in the area of text similarity finding, Braumoeller & Gaines found that EVE2 detected most known textual plagiarism (Braumoeller & Gaines 2001). In the research for the textual similarity, there have been many techniques and mechanisms striving for the better results. In fact the similarity finding in the research article submissions is very tedious task which can be measured only by very effective experiments.

Earlier efforts were made to find the text similarity using string matching algorithms on .rtf format and text documents. Users were allowed to select any of the file formats for comparison. Window's built-in 'Find' functionality is used when exact matches are selected. This is stated to be fast but can miss similarity for reasons like mismatch for any case sensitivity, punctuation mistakes *etc*. Therefore Cerberus worked to find common text clusters which may be useful during the verification and investigation stages of the process,

In another development, Colpaert (Colpaert *et.al.*, 2003) recommended his own algorithm used for approximate matching. The algorithm developed by him is said to be more accurate but much slower than the exact matching algorithm. The algorithm developed for the approximate matching is calculated as a function of the number of two character sequences in two fragments which are there in common. In the large and longer chain of characters in the document, the measures of the similarity shown by this technique in respect of "unrelated paragraphs" were found to be overmatching. Medori, Atwell, Gent and Souter (Medori *et al.*, 2002) carried out some very specialized tests, based on finding possible plagiarism in laboratory practical reports (Medori *et al.*, 2002). This problem is said to differ from the standard plagiarism detection problem due to increased constraints in the way that submissions have to be presented. These include a standard format of 'introduction, method, results, discussion'. However the research submissions are expected to abide by almost same sections but the vocabulary that is commonplace in scientific work generally becomes unusual and unbounded. Hence current engines are argued to be of only limited suitability for finding plagiarism in research papers where infinite variations may occur under any of their sections.

In another noteworthy research paper, results were related to the potential use of text compression techniques (Medori *et al.*, 2002). This does have shortcomings since under this Saxon's project the idea was first tested. However its efficacy was not acknowledged (Saxon 2001). There are number of different variations tried and results found were presented roughly. Most of them are inappropriate due to assumptions that all submissions will be of an equal length. In another example with the Saxon methodology where two plagiarized submissions, when concatenated together, should compress to around same size is also questionable in real practice. The findings suggested that the technique used will work successfully only when the amount of similarity is large however, when the similarity is moderate then this technique will give poor results. So the technique is not suggested for the improved usage because the corpus used to test the compression technique is limited.

3. The Prime Methodologies

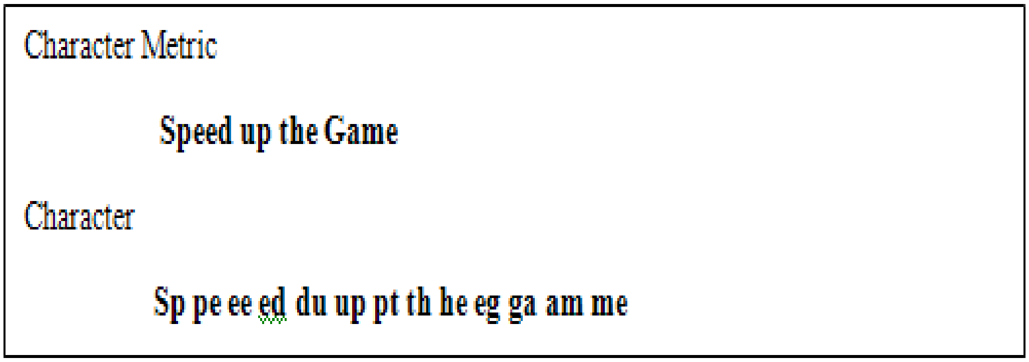
Many different techniques and methodologies have been proposed, tested and suggested by the researchers working in the technical area for finding the textual similarity or in respect of the "textual plagiarism detection". For finding the textual similarity on the local corpus or on the Web (Internet), researchers have proposed and tested different techniques and they got different results and limitations for the same. In the current paper we have analyzed the three most important techniques and methodologies for the textual similarity detection. These are described as below.

3.1. Textual Similarity Detection Using Metrics

The extent case of textual similarity in a document submitted to the corpus can be computed with single metric. In a submission, there may be number of different such metrics associated with it and the similarity could be found in terms of the presence of the different metrics. The vector of metric values representing a submission is known as its fingerprint. In submitted query document, metrics might be related or unrelated. In the practice, the submitted document can be found to have the most relevant or similar fingerprints or dissimilar fingerprints. The most similar fingerprints show that the submitted query document is having closest similarity with the related documents which would mean that every sentence in the submission affected the fingerprint by influencing one of the comparative values. That means common fingerprints have closest similarity.

3.1.1. The Diagram Character Metrics

In this technique, the closest fingerprint for the submitted document is calculated after due normalization. For the normalization process, results for individual metrics can be scaled to range between two chosen values, say 0 and 100, so that each is distributed with the same underlying characteristics. Another part of the methodology is the weighting. In these calculations, the metrics can be multiplied with suitable value and then that metric should be reckoned to arrive at the overall closeness calculations. The Figure 1 given below illustrates the metric calculation for the closest similarity findings.



Character Metric

Speed up the Game

Character

Sp pe ee ed du up pt th he eg ga am me

Figure 1. The Metric Calculation for the Character based Closest Similarity Finding

In practice, the paired metrics also gives a direct measure of the similarity between two submissions. In fact, the normalization process is enforced after the fingerprints generations. Then the paired metrics similarity could be found out to assess the similarity in the submitted documents.

The Figure2 shows the calculation of the fingerprints similarity with the metrics. In this figure, an overall general framework has been shown for the methodologies behind the metric uses and the similarity finding in the submitted document with most related or similar fingerprints calculations.

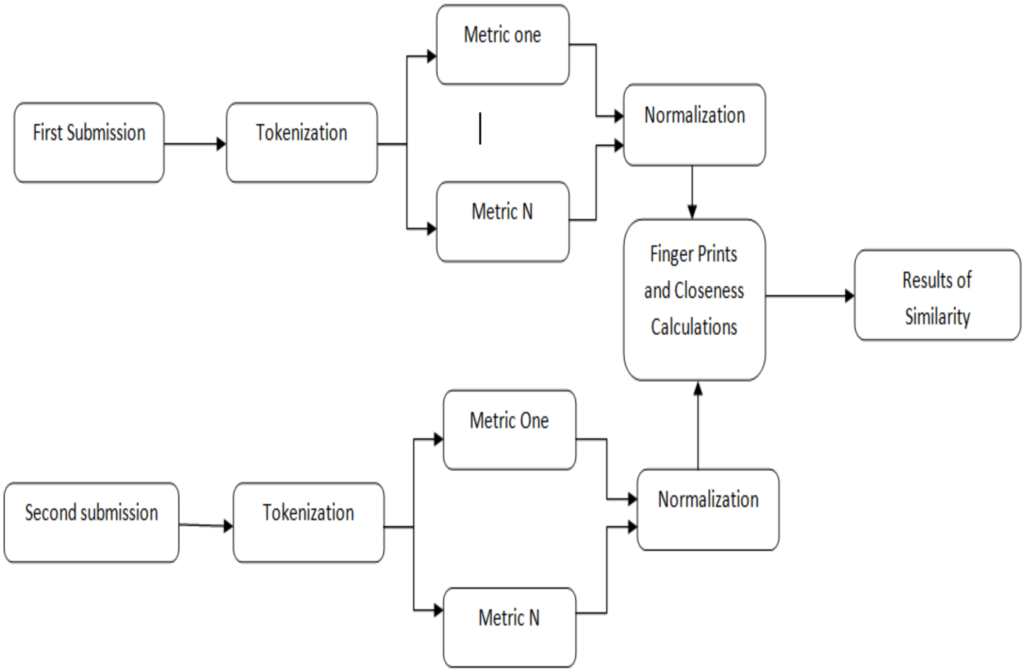


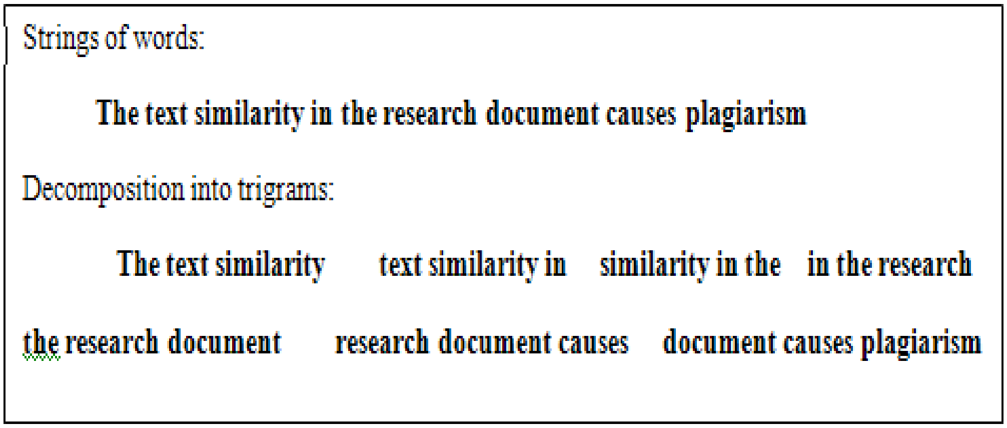
Figure 2. The Overall Working Methodology of Corpus Similarity Calculation with Metric

Source: Thomas Lancaster (2003), Effective and Efficient Plagiarism Detection, School of Computing, Information Systems and Mathematics, South Bank University

3.1.2. Detection and Prevention of Plagiarism based on Trigram Methods

The plagiarism detection and prevention can be implemented in many ways for which the situational parameters affecting the results could be processed by machine or manual way as suited. In the trigram methodology of the plagiarism detection system, the system compares the query document with all and each target document and produces a ranked table of texts with a resemblance measure for each pair. Finally the display of the output could be made side by side with similar passages highlighted. Passages do not have to match exactly. If there is any similarity checked up, after the process is completed, the manual work has to be done, for finding of the exact similarity. Consideration has to be given that even in a case that exact similarity occurs and if some one has given due references, then position must be omitted while concluding the amount of plagiarism. (Lyon *et al.*, 2001)

In the aforesaid methodology, text can be characterized by the set of short sequences of words which they are composed of, typically taken as three-word sequences or trigrams. The Figure 3 given below is an example of the text decompositions in terms of the trigrams. This might be called a fingerprint. Of course, the set of trigrams constitutes a larger text than the original document.



Strings of words:

The text similarity in the research document causes plagiarism

Decomposition into trigrams:

The text similarity text similarity in similarity in the in the research

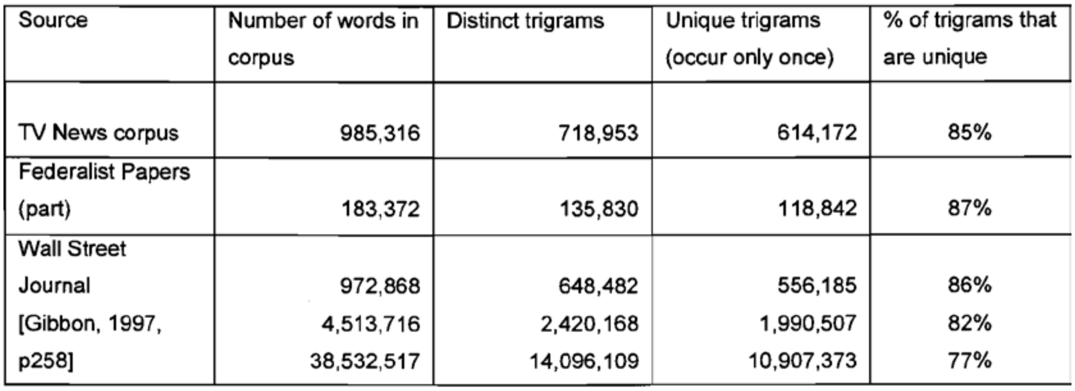
the research document research document causes document causes plagiarism

Figure 3. Text Decomposition into Trigrams

Source: Caroline Lyon, Ruth Barrett and James Malcolm, (2004), A theoretical basis to the automated detection of copying between texts, and its practical implementation in the Ferret Plagiarism and collusion detector, Plagiarism: Prevention, Practice and Policies Conference, St James' Park, Newcastle

In the test cases prepared by the research team using the trigrams for the texts of 1000 – 5000 words, the proportions of matching trigrams is not more than 8%. This is the case even when the same person writes the documents on a similar subject on different occasions. Experiments were carried out on the well-known Federalist Papers, an exhaustively analyzed set of essays like the “Foundation of the American constitution”. In this corpus, the same subjects are addressed repeatedly, and 81 texts were examined. The main purpose of the experiment was to establish a threshold up to which independently written texts might resemble each other (Lyon *et al.*, 2001). Above this threshold, copying or collusion is suspected. The Table 1 given below is the experiment example of the predominance of unique trigrams taken into account of statistics from a TV news corpus, the Federalist papers and the wall Street Journal corpora.

Table 1. Statistics from a TV News Corpus, the Federalist Papers and the Wall Street Journal Corpora



Source	Number of words in corpus	Distinct trigrams	Unique trigrams (occur only once)	% of trigrams that are unique
TV News corpus	985,316	718,953	614,172	85%
Federalist Papers (part)	183,372	135,830	118,842	87%
Wall Street Journal [Gibbon, 1997, p258]	972,868	648,482	556,185	86%
	4,513,716	2,420,168	1,990,507	82%
	38,532,517	14,096,109	10,907,373	77%

Source: Thomas Lancaster (2003), Effective and Efficient Plagiarism Detection, School of Computing, Information Systems and Mathematics, South Bank University

In these experiments, the phenomena of low levels of matching trigrams is seen as the result of the characteristics zipfian (A distribution of probabilities of occurrence that follows

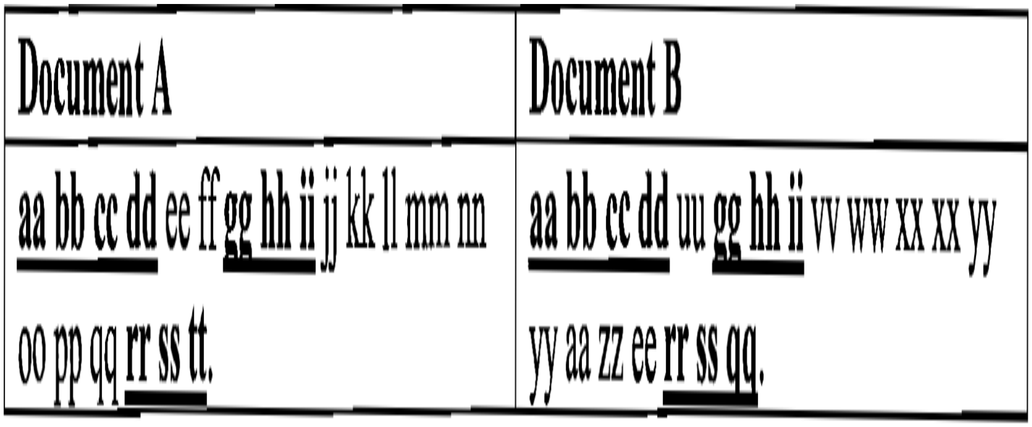
Zipf's law or *The distribution of words is often proportional to a function like $1/n^a$, source Wikipedia – "Zipf's law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. Thus the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc"*) distribution of words in English and other languages. In most of the cases, a small number of words are common, but a significant number of words occur infrequently (Shannon 1951, manning and Schutze 1999). For instance, in the Brown corpus of 1 million words, 40% of the words occur only once (Kupiec 1992). This characteristic is more marked for bigrams and even more pronounced for trigrams. In the trigram methodology, the characteristic of distribution of the trigrams is immediately apparent visually using the Ferret, where matching word sequences are highlighted. When two independent written documents are displayed side by side, there will be scattered highlighted matching word sequence.

3.2. Finding Similarity with the Visualization Method

Finding the textual similarity in the submitted documents is an arduous task. It would be much more beneficial for the publication houses if they could found out the portions of the two sections of the different documents as similar and how similar they are? Some of the new developments in this regard came into existence in early 21st century. For example the Web-based plagiarism detection services present hyperlinks of the line sections of text that are believed to be similar.

Many methodologies have been tested for finding out the plagiarism in the text. One such method is fragments and visualization method. In this method, a document can be considered as a number of shorter fragments, each containing equal number of words. The fragments of the two submissions that are believed to contain similarity may map on the re-application of simple metrics. The metrics could be said to be applied at fragmentary granularity as opposed to the gross granularity at which the regular similarity measurement engines work. Hence for the submissions one can find the similar fragments. The Table 2 below shows how the small fragments of the documents A and B have the similarity of the highlighted fragments.

Table 2. Shows the Small Fragments of the Document A and Document B as well the similar Fragments

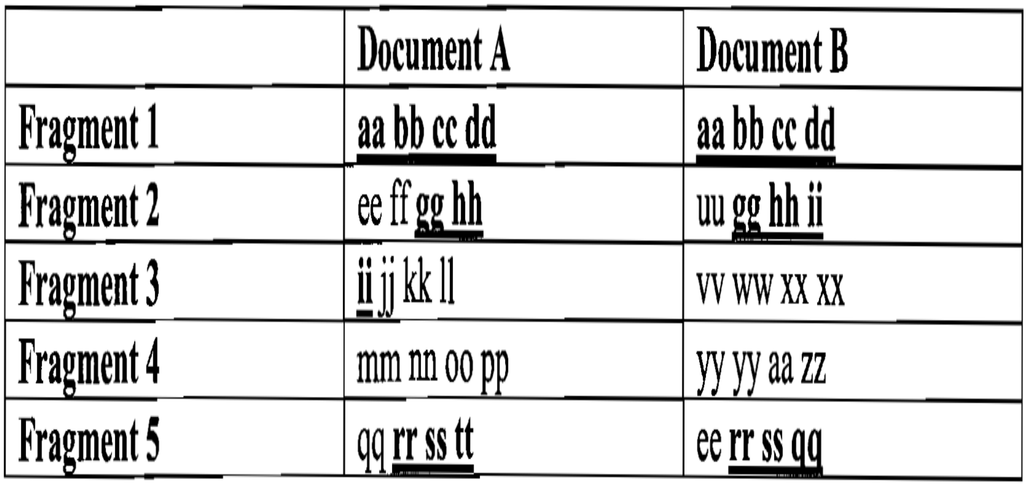


Document A	Document B
aa bb cc dd ee ff gg hh ii jj kk ll mm nn oo pp qq rr ss tt.	aa bb cc dd uu gg hh ii vv ww xx xx yy yy aa zz ee rr ss qq.

Source: Thomas Lancaster (2003), Effective and Efficient Plagiarism Detection, School of Computing, Information Systems and Mathematics, South Bank University

In the Table 3, an example of the two documents similarity has been demonstrated for the small fragments. The comparisons show a four word run of exact similarity at the beginning of the both documents, a three word run of exact similarity inside both documents and a three word run of arranged similarity at the end of both documents.

Table 3. The Small Fragments Shows the Document A and Document B of each Four Words



	Document A	Document B
Fragment 1	**aa bb cc dd**	**aa bb cc dd**
Fragment 2	ee ff **gg hh**	uu **gg hh ii**
Fragment 3	**ii** jj kk ll	vv ww xx xx
Fragment 4	mm nn oo pp	yy yy aa zz
Fragment 5	qq **rr ss tt**	ee **rr ss qq**

Source: Thomas Lancaster (2003), Effective and Efficient Plagiarism Detection, School of Computing, Information Systems and Mathematics, South Bank University.

The examples of the documents similarity given as these are the ones with high valued cells. Where some fragments of the document A exactly match with the fragments of the document B. In the methodology of the fragments similarity on the basis of matches, one then calculates the percentage of the similarity in the two documents. More the similarity of the fragments, more is the percentage of the similarity in the submitted document.

In the similarity measure of the small fragments, the documents can be considered to be set of fragments and their interception matrix may prove useful for identifying the fragments of two documents that are most similar. The documents submitted by the user may be large or big in text size so the fragments of the documents may be checked in the one, two or three word fragments but if the fragments gap is considered for one, then alternate fragments must be considered. If the fragments gap is chosen two, then every second fragments of the documents must be considered to maintain the sequence of the whole documents. A representation of the first submitted document can now be compared with the second document. For accurate comparison the second document is extracted using the same fragment size and gap size as the first. This comparison of two fragments gives a fragmentary interception, a representation of the contents of both fragments. In this process, it calculates the fragmentary interception similarity score. A similarity score as localized to the fragmentary interception, is scaled between 0 and 100, with 0 representing no similarity and 100 representing the identity.

For finding the similarity score in the submitted document, the visualization of the whole similarity area of the document can be plotted by the pixel visualization. In the fragments visualization, gray level information can be used to display the extent of similarity within fragments. For each cell within the fragmentary intercept similarity score matrix, it can be mapped onto a grayscale value, with a fragmentary similarity score of 0 corresponding to white and a score of 100 corresponding to black. The grayscale values can be plotted onto pixels onto a graphic plane known as similarity visualization. In this process anyone can evaluate the similarity area and can find out the similar text in the submitted documents. In the pixels joining together in this process, it makes some areas as prominent. Such areas are known as similarity intersections. In the Figure 4 the visualization of the examples taken in the documents may be reckoned to be plagiarized. The visualization has been produced with a fragment size of 200 words and a fragment gap of ten words, where documents submitted was on the y-axis and the web resources on the x-axis.

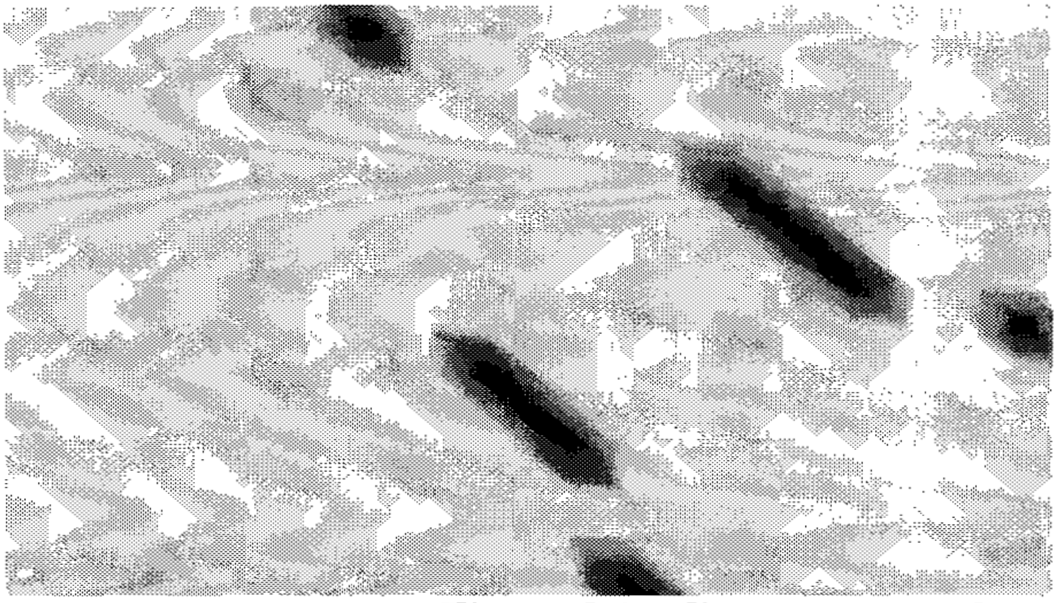


Figure 4. The Example of the Similarity Visualization

Source: Thomas Lancaster (2003), Effective and Efficient Plagiarism Detection, School of Computing, Information Systems and Mathematics, South Bank University

In the above visualization, the intense black area of the figure are plagiarized portions and the other part of white areas are rest of the document. So the visualization can be seen in the example for the finding similarity in the submitted text.

4. Results Analysis

For the text similarity finding in the documents submitted for the publications and other academic uses, much suitable search is required for finding the exact similarity. The technique or methodology used is the metrics based on the fragments of the words. There are no such facilities for finding exact similarity. The use of word count techniques for the large set of documents and larger corpus will be very tedious and time consuming process. The complexity of the process will also be too high because the process will count each and every word for the similarity check and then generate the performance results.

In the metric calculations, the submitted document compares with the similar corpus documents which are also fragmented into metrics and then compared with the submitted document. After the similarity check of the word in the documents, there is a process of normalization which is used for some calculation of the percentage similarity and mathematical measures for the results that can say what are the exact positions of the similarity of the text. In this whole process, there is less confidence for the results that can say the exact similarity of the documents with the target similar documents.

The trigram technique is primarily directed to finding the text similarity on the word level after due removal of stop words and performing the process of stemming. It finally uses string matching mechanism for the similarity finding in the submitted document. Again, selection is made only of 3 words and follows the trigram mechanism. As a consequence, enumerable numbers of matches are to be performed and thereafter further metrices formulations and comparisons are performed making the computation process very tedious and time consuming. Again in real field, the user never agrees on the found plagiarism report simply based on the count of word matches, since no direct evidence of plagiarism is seen by a common man. The trigram technique used is valid for the documents which are related to corpus or are small in size. When used for the larger size of the documents and on the Web based search, it becomes very complex. When the document is large and searching is performed on the web, then matching with the three word sequence becomes unconverging process for the machine.

In another methodology involving use of fragments matrix intersection and visualization, the mechanism is flexible for the word count. The user can decide the word fragments and then set the count of the word fragments either 3 or 4 or any other. The similarity check on the fragments chosen is performed and then the area of localization is set where the text similarities are found and have plagiarism. The similarity of the corpus documents can be found as well as on the Web. But in the case of Web, the similarity finding mechanism of the techniques will be very complex because on the Web there is huge set of documents available. After the similarity finding of the text, the fragments are mapped with the pixel values to obtain the graphical image of the data towards visualization on the x-y plane, wherein the y axis holds the set of query document and x-axis holds the documents existing on the internet. The image has black and white portions scaled from 0 to 100, with blackish part being near to 100 and whitish part being near to 0. Through this visualization process, the user can find out the similar parts of the text and then manual work has to be done to find the exact similarity and percentage similarity of the submitted document with the documents on the web. Thus the exact similarity finding is not possible in these recent techniques but inexact matches may be worked out satisfactorily.

5. Conclusion

The recent research and developments in the area of finding textual similarity of the submitted document with those in a corpus is advancing day by day. Some of the professional setups in these recent years are providing the services to the universities and research organizations for reporting the textual similarity of their submitted documents. The technologies and methodologies used in these developments for the similarity findings are mostly on the word counts. The word counts in the submitted documents with the similar corpus documents are used to provide the similarity index for the given text. Most of the cases are based on the corpus or local repository based resource documents. Web based repository has also been tried. However in both of these cases, the results found are only a coarse value of plagiarism and are not satisfactory.

In the present paper, the most appropriate and advanced mechanisms of the textual similarity plagiarism detection techniques have been described and comparisons have been made for live problems of the academia and research organizations. The real problem of plagiarism being faced by these academia and research organization is very different in condition described above. Three popular techniques have been analyzed above which are useful only for the small local repositories and corpus. All These deal with the small size of the documents submitted for the similarity checkups which are performed on the small corpus to generate the word count index as results of the percentage plagiarism. In the real world, the problem is very vast and unbounded in nature. The user writes the documents, sometimes including copied portions from the Internet but the huge data from the Internet needs to be compared to find similarity. In the mechanism of the word count on the basis of trigram or visualization process, when the data is huge and chunks of the documents are plagiarized or paraphrasing is used, then the stated mechanisms, will fail and the complexity of the process will become too high. In the aforesaid mechanisms, each and every 3 word (trigram) will be counted and then the similarity matching will be performed. Thus its complexity will become too high and when the resources of the copied documents are different then it will be unable to locate the similarity with those documents. So it calls for the new mechanisms to be developed for the Web based resources repository matching which may be performed on the sentences basis. The similarity check may thus need to be addressed for chunks of data with local repository as well as those found relevant on the Internet.

References

[1] M. Austin and L. Brown, “Internet Plagiarism: Developing Strategies to Curb Student Academic Dishonesty”, The Internet and Higher Education, vol. **2**, no. 1, **(1999)**, pp. 21-33.

[2] J. Buckell, “Plagiarism Tracked at 8 Per Cent”, Australian IT, 11/09/02, Available online at http://australianit.news.com.au/articles/0,7204,5071781%5e15334%5e%5enbv%5e15306-15317,00.html, **(2002)** November 9.

[3] P. Ahlgren, B. Jarneving and R. Rousseau, “Requirements for a cocitation similarity measure, with special reference to Pearson’s correlation coefficient”, Journal of the American Society for Information Science and Technology, vol. 54, no. 6, pp. 550-560, **(2003)**.

[4] P. Ahlgren and C. Colliander, “Document–document similarity approaches and science mapping: experimental comparison of five approaches”, Journal of Informetrics, vol. 3, no. 1, **(2009a)**, pp. 49-63.

[5] P. Ahlgren and C. Colliander, “Textual content, cited references, similarity order, and clustering: an experimental study in the context of science mapping”, Proceedings of the 12th International Conference on Scientometrics and Informetrics, Rio de Janeiro, vol. 2, **(2009b)**, pp. 862–873.

[6] P. Ahlgren and B. Jarneving, “Bibliographic coupling, common abstract stems and clustering: A comparison of two document–document similarity approaches in the context of science mapping”, Scientometrics, vol. 76, no. 2, pp. 273-290.

[7] K. W. Boyack, R. Klavans and K. Borner, “Mapping the backbone of science”, Scientometrics, vol. 64, no. 3, **(2005)**, pp. 351-374.

[8] K. W. Boyack, D. Newman, R. J. Duhon, R. Klavans, M. Patek and J. R. Biberstine, “Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches”, PLoS One, Article Number: e18029, vol. 6, no. 3, **(2011)**.

[9] K. W. Boyack and R. Klavans, “Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?”, Journal of the American Society for Information Science and Technology, vol. 61, no. 12, **(2010)**, pp. 2389-2404.

[10] COLPAERT, Jozef and Wilfried DECOO (2003): “String-matching algorithms as Web Services for answer evaluation, dictation correction and plagiarism detection”, EuroCall, Limerick, Ireland, 5 September.

[11] Cristian Colliander and Per Ahlgren, “Experimental comparison of first and secod-order similarities in a scientometric context”, Scientometric Springer Publications, vol. 90, **(2011)**, pp. 675-685.

[12] F. Culwin and T. Lancaster, “A Descriptive Taxonomy of Student Plagiarism”, unpublished, available from South Bank University, London, UK, **(2000a)**.

[13] F. Culwin and T. Lancaster, “A Review of Electronic Services for Plagiarism Detection in Student Submissions”, Proceedings of 1st LTSN-ICS Conference, Edinburgh, **(2000b)**, pp. 54-61.

[14] F. Culwin and T. Lancaster, Plagiarism Issues for Higher Education, Vine 123, available from LITC, South Bank University, London, **(2001a)**.

[15] F. Culwin and T. Lancaster, “Plagiarism Prevention, Deterrence & Detection, available online at http://www.ilt.ac.uk/resources/Culwin-Lancaster.htm, **(2001b)**.

[16] L. Egghe, “New relations between similarity measures for vectors based on vector norms”, Journal of the American Society for Information Science and Technology, vol. 60, no. 2, **(2009)**, pp. 232-239.

[17] L. Egghe, “Good properties of similarity measures and their complementarity”, Journal of the American Society for Information Science and Technology, vol. 61, no. 10, **(2010a)**, pp. 2151-2160.

[18] L. Egghe, “On the relation between the association strength and other similarity measures”, Journal of the American Society for Information Science and Technology, vol. 61, no. 7, **(2010b)**, pp. 1502-1504.

[19] L. Egghe and L. Leydesdorff, “The relation between Pearson’s correlation coefficient r and Salton’s cosine measure”, Journal of the American Society for Information Science and Technology, vol. 60, no. 5, **(2009)**, pp. 1027-1036.

[20] L. Egghe and R. Rousseau, “Classical retrieval and overlap measures satisfy the requirements for rankings based on a Lorenz curve”, Information Processing & Management, vol. 42, no. 1, **(2006)**, pp. 106-120.

[21] M. Gmur, “Co-citation analysis and the search for invisible colleges: A methodological evaluation”, Scientometrics, vol. 57, no. 1, **(2003)**, pp. 27-57.

[22] L. Hamers, Y. Hemeryck, G. Herweyers, M. Janssen, H. Keters and R. Rousseau, “Similarity measures in scientometric research- The Jaccard index versus Salton cosine formula”, Information Processing & Management, vol. 25, no. 3, **(1989)**, pp. 315-318.

[23] R. Klavans and K. W. Boyack, “Identifying a better measure of relatedness for mapping science”, Journal of the American Society for Information Science and Technology, vol. 57, no. 2, **(2006),** pp. 251-263.

[24] A. Lathrop and K. Foss, “Student Cheating and Plagiarism in the Internet Era – A Wake Up Call”, Published by Libraries Unlimited Inc, **(2000)**.

[25] L. Leydesdorff, “On the normalization and visualization of author co-citation data: Salton’s cosine versus the Jaccard index”, Journal of the American Society for Information Science and Technology, vol. 59, no. 1, **(2008)**, pp. 77-85.

[26] T. Luukkonen, R. J. W. Tijssen, O. Persson and G. Sivertsen, “The measurement of international scientific collaboration”, Scientometrics, vol. 28, no. 1, **(1993)**, pp. 15-36.

[27] H. Maurer, F. Kappe and B. Zaka, “Plagiarism- A survey”, Journal of Universal Computer Science, vol. 12, no. 8, **(2006)**, pp. 1050-1084.

[28] J. Kupiec, “Robust part-of-speech tagging using a Hidden Markov Model”, Computer Speech and Language, vol. 6, **(1992)**, pp. 225-242.

[29] C. Manning and H. Schutze, Foundations of Statistical natural Language Processing, MIT Press, **(1999)**.

[30] C. Shannon, Prediction and Entropy of printed English, in Shannon, C.E Collected Papers, Sloane and Wyner (eds). IEEE Press, **(1993)**.

[31] P. Samuelson, “Self Plagiarism or Fair Use, Communications of the ACM, August, vol. 37, no. 8, **(1994)**, pp. 21-25.

[32] H. P. F. Peters and A. F. J. Van Raan, “Co-word-based science maps of chemical-engineering”, Part 1: Representations by direct multidimensional-scaling. Research Policy, vol. 22, no. 1, **(1993)**, pp. 23-45.

[33] H. Maurer, H. Krottmaier and H. Dreher, “Important Aspects of Digital Libraries”, International Conference of Digital Libraries, New Delhi, **(2006)**, December 5-8.

[34] N. J. van Eck and L. Waltman, “How to normalize cooccurrence data? An analysis of some wellknown similarity measures”, Journal of the American Society for Information Science and Technology, vol. 60, no. 8, **(2009)**, pp. 1635-1651.

Authors

Ranjeet Kumar has been working in Patent Referral Centre which is having an Anti Plagiarism Cell in Indian Institute of Information Technology Allahabad since 2007. He has published 10 research papers in the field of IPR. He has been vetting the originality of expressions in MTech (IT) and Ph.D thesis works of the Institute. The plagiarisms check up and novelty assessments in the IT patents are his main activity.

Prof. R.C.Tripathi is Incharge of the Patent Referral Centre which is having an Anti Plagiarism Cell in Indian Institute of Information Technology Allahabad. He has been Dean (R&D) since 2007 to 2011 and since 2007 he has been promoting IPR culture in IIIT Allahabad. Earlier he won the Government of Netherlands Fellowship in 1975 and has worked for 30 years in different capacities to promote R&D in the country as an officer in the Ministry of Communications and IT Govt. of India where he also headed the “Patent and IPR Division” and looked after enhancing IPR portfolio of the 11 laboratories in the country setup by the said Ministry.