A Novel Approach for Plagiarism Detection in English Text

Shivani
Department of Computer Science
Punjabi University,
Patiala

Vishal Goyal, PhD
Department of Computer Science
Punjabi University,
Patiala

ABSTRACT

Digitalization provides text easily available on web interrelated to several academic areas. So it becomes a serious problem for academic enterprises or institutes. This paper presents Plagiarism detection system for the English language. Digital World provides text easily available on web interrelated to several academic areas. So it becomes a serious problem for academic enterprises or institutes. PD means to detect the text being copied from original sources through websites, books, journals, previously published papers, online search engines, etc. this paper have presented the development of a web-based PD system to discover the similarity in English written text only. This paper is going to discuss textual based PD on an exact string matching technique through the DB and the web. The proposed system has presented concerning principal behind the system. The proposed system supports three steps: first is Pre-processing where the splitting of the input string to sentences and stop words are removed. Second is the process of sentence searching through DB and the web. Once plagiarized sentence is already there in DB then sentence directly retrieved from DB with stored URL. If searching of the sentence is not found there in DB, then plagiarized sentence is searching throughout the web ("GOOGLE") starts for both semantic and syntactic by using Cosine Similarity Approach. After Web search plagiarized sentence is stored in the DB. Thirdly, similarity analysis is performed for detail description about all plagiarized sentences with the URL (source address). As a result, the proposed system displays plagiarized sentences with the original source's URL and percentage of Plagiarism within the input string.

Keywords

NLP, Plagiarism Detection, Textual Similarity, Exact string matching scheme, Results analysis by comparison with other tools.

1. INTRODUCTION

NLP (Natural Language Processing) is the branch of computer science and computational Linguistics work as the interface between human understandable languages and the computer system [1]. Natural language processing resources to build the software that is able to comprehend generate and study the human language that human use naturally [2]. NLP is a function of a computer system that extracts NLP input and produces NLP output. NLP provides search capabilities to get better and effective result by utilizing its feature and to capture the result, whether the text is rewritten form of some former source text or not [3].

1.1 Plagiarism Detection (PD)

"Plagiarism is the act of taking the writings of another person and passing them off as one's own. The fraudulence is closely related to forgery and piracy-practices generally in violation of copyright laws". Encyclopedia Britannica [4].Plagiarism stands for "Copying" and Detection stands for "Revealing". So Plagiarism Detection is to detect or reveal the text or document that has been copied (textual similarity). PD has become a serious offense these days, especially in the academic work with the usage of progression in the technology. The computer is that electronic device that is worn for this crime like any other crimes similar to computer hacking, phishing, spamming, etc. PD for NLP was first developed after 1990's. The first PD system was MCQ (Multiple Choice Questions) detection system. But now in today's world, everything is just one click apart. By just clicking one button, it's become easier to copy or steal the text of another person, without citation. By this deserving candidate doesn't get the credit for their work [5]. There are many ways to accomplish the duplicated text from other sources. So Plagiarism is classified into the following types:

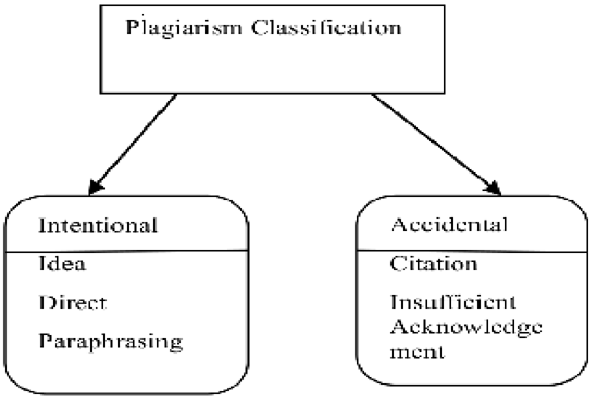

Figure 1 Classification of Plagiarism

Intentional Plagiarism: It is a kind of plagiarism where writer already knows the original source of plagiarism (online source, pre-research papers, books, etc.) but intentionally not given the authentic credit to the source (intelligence dishonesty). Idea plagiarism has borrowed the words (cliff notes) of others without credit. Direct plagiarism is when a person steals the text directly without citation. Paraphrasing has rearranged the work of another with the same significance. The patch is cut and pastes the parts of text from many sources [6].

Accidental Plagiarism: It is a kind of plagiarism where writer fails to give credit to original source. The reason could be poor or careless paraphrasing, less knowledge about citation or forget footnotes. Citation Plagiarism is when the writer forgets to cite. Insufficient acknowledgment is when a flawed citation is given. Mosaic is when due to lack of knowledge; writer ignores to giving a quotation[6].

Both above plagiarism can be detected either manual (human) or automatic (machine). But automatic PD is an easy way to detect because it takes less time, efforts and need not memorize the content. PD exists in many forms: textual PD, source code PD, citation based PD, PD on the basis of author name with the title of work based. But this paper mainly discussed textual based PD with their valid approaches and schemes.

1.2 Textual Plagiarism Detection Technique [7]

This is a form of PD that essentially impacts on "written text" only. Textual Plagiarism Detection is a most common type of Plagiarism that basically belongs to the academic enterprise in the student's assignment, paper writing, document submission, etc. Plagiarism Detection has become a standard in education institutes, where no. of resources increasing day by day. It became difficult to detect the plagiarism with limited resources because manual textual detection needs more effort, memory, and span of time than to automatic detection. Automatic detection is necessary to check plagiarism. With this benefit of automatic detection, a vast amount of data gets easily checked by using the internet resources or search engines. Automatic textual checking is considered as an operation that applies to the altered text in the NLP field.

The Textual Plagiarism Detection Technique has two considerable approaches:

Exterior Approach: The plagiarized sentence is considered as replicated under exterior approach when plagiarism is checked in the between textual data by comparing the plagiarized document with the original document. This approach elevates to find the plagiarized text with the source of copied and % of plagiarism occur during the search of exact string matching.

Interior Approach: The plagiarized sentence is considered as replicated under interior approach when writing style is different in a single document. This implies when someone writes some part of a document by itself and another part by another person. Interior Detection has no dependency on Exterior Detection for similarity checking, but this approach can't give the actual source of plagiarism.

The textual similarity is checked by considering words that contain useful information. There are so many forms to find similarity between suspicious sentence and the original sentence of textual data. Some of them are discussed below [8]:

Word Similarity: this similarity is used to detect the similar words into the sentences after stop words are removed. Let 'p' and 'q' are two sentences of words. The similarity of common words between them is calculated by using a formula:

$$similarity(p,q) = 2 * \left({sw(p,q)}\big/{(length(p) + length(q))} \right)$$

The "sw(p, q)" is no. of similar words between the sentences and length (p), length (q) are the length of words in sentence p and q respectively.

Word order similarity: this similarity is used to detect the sequence similarity between the sentences. Here this needs to find the vector word order for each sentence. Let V_p and V_q are the corresponding vector word order for both sentences p, q. Word order similarity is calculated by using a formula:

$$WoSimilarity(p,q) = 1 - \frac{||V\text{p} - \text{Vq}||}{||Vp + Vq||}$$

Word sentence similarity: this similarity is used to detect the maximum similarity between the word 'a' and words in sentences 'A'. This similarity is calculated by using a formula:

$$WsSimilarity(a,A) = maximum[similar(a,Aw)]|Aw \in A$$

WsSimilarity is Word sentence Similarity, (a, Aw) are words and similar (a, Aw) are word similarity between 'a' and 'Aw'.

Word Semantic Similarity: this similarity is used to detect the semantic similarity between the sentences 'p' and 'q'. The WSSimilarity is calculated by using the following formula:

$$WSSimilar(p,q) = \frac{\sum WsSimilarity(aw1,q) + \sum WsSimilarity(aw2,p)}{|p| + |q|}$$

| A | is the no. in the sentence A.

Sentence Similarity: this similarity is used to detect the similarity between the sentences. This similarity has described a no. between 0 and 1. 0 stands for not a similar sentence and 1 stands similar sentence. The similarity is more, as no. is increasing from 0 to 1. This Sentence Similarity measure is calculated by using a formula:

$$SSimilarity(p,q) = \alpha 1 * similarity(p,q) + \alpha 2 * WoSimilarity(p,q) + WSSimilarity(p,q)$$

SSimilarity is Sentence Similarity and α1, α2, sα3 are constants and α1+α2+α3=1

This paper basically concentrates on textual based plagiarism detection by utilizing Exterior Detection Approach under both the web (Google) and offline (database) checking. The similarity measures in exterior approach are cosine similarity. In the rest of the paper, that examined about a review of different plagiarism detection tools, scope, and methodology of the proposed system by means of comparison with other existing tools to evaluate results. At the end of the paper, a conclusion with the outlook of the system is maintained.

2. LITERATURE SURVEY

Many types of research on Online Plagiarism Detection have already been proposed. For Online Plagiarism Detection in Natural Language text (English), that has a survey in detail on appeal literature. During the survey, gets a chance to learn about various perspectives and approaches that already been developed for Plagiarism Detection [9].

Shivakumar and Garcia-Molina [10] developed SCAM System. This system basically used to find out the comparison between two documents on the basis of a set of words by using word analysis.

Asim, Ali, Hussam and Vaclav [11] proposed the approach for comparison of five different kinds of software that are used for textual plagiarism detection. The comparison is done on the basis of their features and performance. They find there is still no software that detects or prove 100% plagiarism because each software and tools has advantages and limitations.

Antonio, Hong VA and Rynson WH [12] developed CHECK System to identify the identical documents that have the same domain. Example: Computer science.

Richa, Puneet and K. Nithyanandam [13] highlight the NLP approach where all those software's that are freely available online. They also highlight all the free software with the URL and their method of use with pros and cons of the software.

Monostori, Arkady and Heinz [14] built MatchDetectReveal System proposed the exact string matching by using the algorithm. Even after requiring more space and time, accuracy is the best part of this System.

Yaakov, Aharon and Natan [15] developed software for simple plagiarism detection and also built a corpus consists of 10,100 papers of CS in English. They chose the baseline method to identify identical paper. This software shows Plagiarism on the basis of similarity in abstract and references of the papers.

There are numerous online tools accessible already in the market that is considered under textual matching. This paper is going to examine the working of some online tools by utilizing their website address and later contrasted these tools and our system. Some existing tools are as follows:

Duplichecker is an online web-enabled tool, where the user has the facility of direct copy paste detection or upload facility. After clicking on "search button", this tool compares content with online sources. This instrument gives results in an HTML page structure. This tool is anything but difficult to utilize yet limited no. of search available.

PlagScan is an online textual checking instrument which is specially intended for academic purpose where duplicated text is compared with books, journals, published papers etc. Web similarity checking is also accessible. Where the user can enter the text or upload the file. The report is prepared after clicking "check" button. But this tool has a composite boundary for detection with multiple checking.

Plagiarisma is an online detection tool that provides results by checking similarity with already stored and also with Google books and scholar. This tool also supports multiple languages, but membership charge needs to pay only for restricted no. of options to check the tracking about no. of detection per day with the restriction on the address of the system. This system works well only for exact matching not for paraphrased text.

Plagiarism software is a kind of web detection tool, where the user can copy the text in the search area and click on "check" button. Plagiarized sentences are compared with online sources and provide the results. The similarity is doing a line by line procedure without any PDF file checking facility.

Plagchecker is a web plagiarism checking tool, where the user can paste the content into giving box and afterward subsequent to entered the security code (already given) then click on the green button and compare written text online line by line. Website inspection and cross-examine features are utilized by this tool.

Urkund is a web plagiarism detection system that checks documents by upload specifically to the website, by sending email through the instructor / guide. Urkund works under email sending and receiving.

3. SCOPE AND METHODOLOGY OF PROPOSED SYSTEM

Very Recently, For Plagiarism Detection, NLP approaches are budding in favor of System Efficiency. So it is necessary to elaborate the scope of our work. The scope of this paper is to discuss the development of Web-based PD System that efforts proficient English has written plain text or input string only and consider that written text is mono-lingual text (copied form of English language only) with both short fragmented and long fragmented lengths. So the proposed system mainly works well for textual based PD with exact string matching. The proposed methodology is implemented in Web Environment. The algorithm has been implemented by using language PHP on XAMPP Platform and individual plagiarized sentences saved by using Microsoft structured query language database. The Methodology that supports proposed system incorporates subsequent steps for the retrieval of exact string matching:

3.1 Pre-processing step

This task is a fundamental task to prepare a system for the basic steps of stop word removal, filtering, substantial approach, etc. First of all, the user enters an input string that consists of no. of sentences to checking the similarity among them. To splitting the input string into individual sentences. On the basis of NLP task, stop words removal is applied for filtering of unnecessary words.

3.2 Processing of plagiarized sentence searching

The similarity for each individual sentence is checked as follows:

3.2.1 Similarity retrieval through database

First of all, sentence searching starts with the database. The user can enter a sentence into textbox area. If that particular sentence exists there in the database with exact matching. Then retrieval of that particular sentence is through the database along with the corresponding URL stored in the database.

3.3 Database Design

Table 1. Table Carries the Design of the Database with their fields with their types and Description

Fields with Types	**Description**
SentenceId(int)	The Unique identification number is given for each sentence enters for plagiarism detection as primary key and foreign key in other related tables.
SearchSentence(text)	Search for plagiarism Sentence stored directly in the database
SentenceURL(varchar)	It contains the Web address of above Plagiarized Sentence
TextFull(long text)	It contains the whole body text of a URL, that is specified as plagiarized

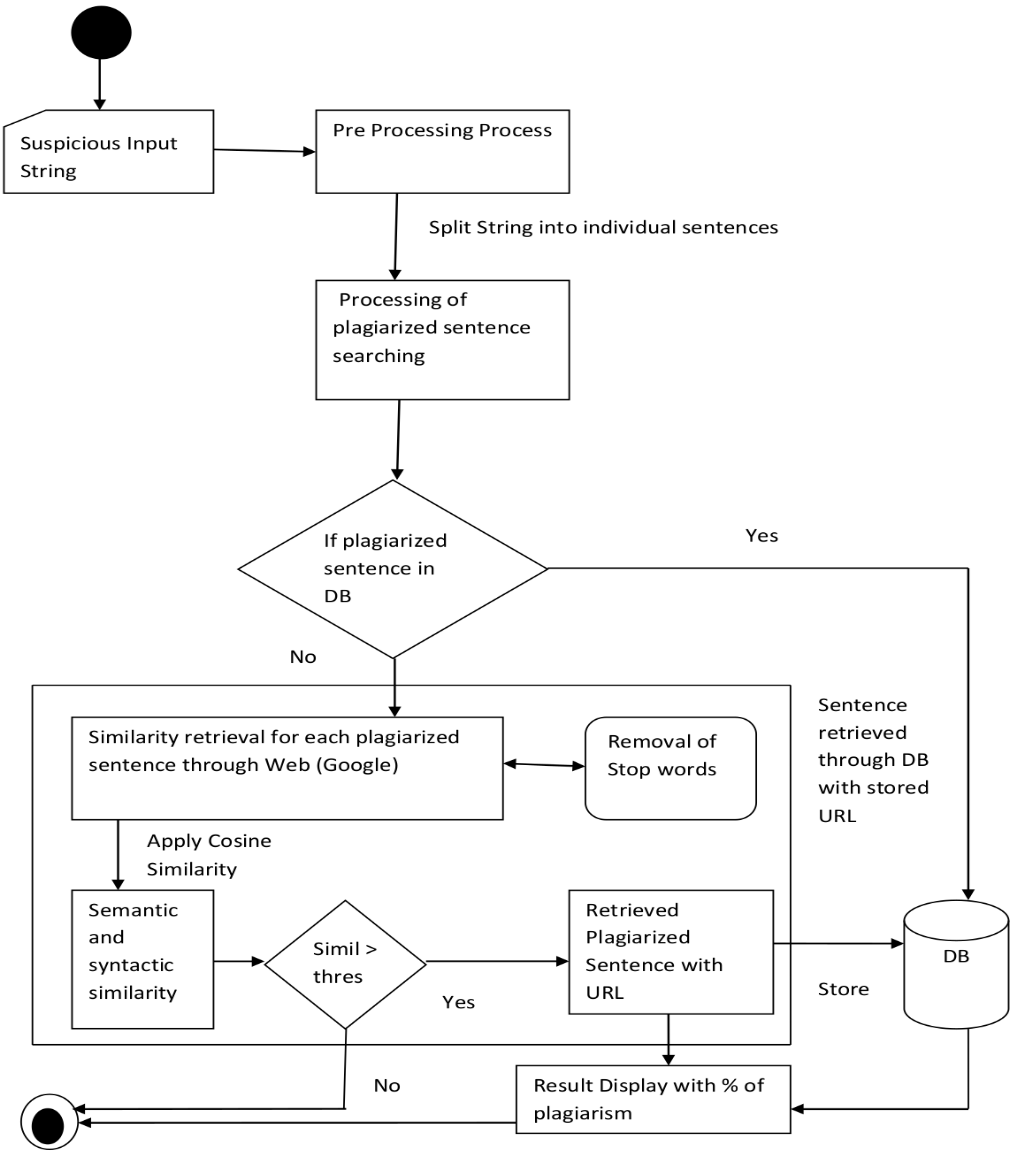
Suspicious Input String
Pre Processing Process
Split String into individual sentences
Processing of plagiarized sentence searching
If plagiarized sentence in DB
Yes
No
Sentence retrieved through DB with stored URL
Similarity retrieval for each plagiarized sentence through Web (Google)
Removal of Stop words
Apply Cosine Similarity
Semantic and syntactic similarity
Simil > thres
Yes
Retrieved Plagiarized Sentence with URL
DB
Store
No
Result Display with % of plagiarism

Figure 2: Flow Chart of Proposed System

2.2.2 Similarity retrieval through the web

If that particular sentence is not found in the database, then that sentence is submitted to web search engine “Google” for the retrieval of the original source of a particular sentence. This web-based retrieval is working as follows:

Before sentence that submitted to the web (Google Search), stop words (source: http://www.ranks.nl/stopwords) are removed from the sentences along with the abbreviations so that efficiency of a search can be increased. Stop words are meaningless words that present in between the sentences; these words need to be removed for better understanding and accurate results (articles, prepositions, abbreviations, conjunctions etc). For example: “a”, “an”, “the”, “at”, “after”, “before” etc. After removal of these unnecessary words, resultant sentences are searched by using Google search engine and the top URL of Google for the relevant sentence is considered. If the sentence’s exact copy is matched on that particular URL, then content within particular URL is retrieved and that plagiarized sentence with URL is stored in the database.During the similarity retrieval through Google Search Engine, stop words removal is necessary to perform. Because maybe after removal of these words, plagiarized sentence will be able to find the similarity to increase the efficiency of the system. Cosine similarity is applied for every sentence here.The Enhancement of PD by machine learning approach is based on both semantic and syntactic text. The MLT considered as a subdivision of CS which is used to encapsulate the data. The MLT composed no. of methods/resources of technique with the algorithm. The MLT is the automated learning technique which is the DB based and supervised learning based where the level of plagiarism is detected for the decision making. Detection of plagiarism between the two documents is checked on the basis of following two levels:

Semantic Similarity: SES is the important concept in the field of NLP where SES clarified the similarity within the two terms of the given input string. SES plays a considerable role in the field of textual similarity within the words of the sentences. When the words within the documents are interchangeable that is synonyms. Semantic is used to identify the expressive meaning within the terms of the documents. This similarity is essential when someone uses another person’s text by replacing the content of words with corresponding similar words.

Syntactic Similarity: SYS is also another important concept of NLP where similarity works on the structure of sentences. If the structural relationship between the two sentences of words is homogeneous, then the syntactic similarity is working. SYS similarity of detection is where words of sentences are collaborating together in a similar manner with the grammatical structure of sentences in the practice of close syntax. The texts for plagiarism may be picked from different sources are combined together in the same procedure. Cosine similarity is applied for every sentence of the input string for similarity detection.

3.4 Similarity retrieval within the input string with % of plagiarism

After the above process repeated for all sentences, the similarity for writing text is retrieved in the form of individual sentences display with their valid URLs. Once the whole string is searched, the result is displayed with the % of plagiarism within the written text or enters text.

Percentage of Plagiarism for complete input string is

Considered by using following formula and its range are between 0 and 1.

$$\%\ \boldsymbol{of\ plagiarism} = \frac{\boldsymbol{no.of\ plagiarism\ sentences}\Big/\boldsymbol{search\ (a)} * \mathbf{100}}{\boldsymbol{Total\ no.sentences}\Big/\boldsymbol{search\ (b)}}$$

% of Plagiarism is 100%, when Number of plagiarism Sentences/search is equal to total no. of sentences taken /search. % of Plagiarism is 0%, when Number of plagiarism Sentences/search is totally different from total no. of sentences taken /search.

$$\boldsymbol{the\ time\ complexity\ is\ O\ (a * b)}$$

4. RESULT AND ANALYSIS

To prepare results of this proposed system by collecting Sets consisting of different text from different places related to different articles. The proposed system is tested for comparison with other freely available tools online opposite by using the similar sets of texts. The proposed system showed results by comparing with Number of sets of text with their percentage of plagiarism within the text. These sets are labeled as P, Q, R, S, T, U.

Table2. Accuracy comparison of different tools

Software Tools /Sets of text	**Proposed system**	**Plagiarism checker**	**Plagscan**	**Plagiarismsoftware**
P	89.79	94	87.4	91
Q	73.58	77	62.1	79
R	86.95	93	97.1	95
S	48.3	42	43.8	52
T	Unique	Unique	Unique	Unique
U	96.07	78	73.1	74
V	61.53	58	51.5	58

The Set P consisted of 50 sentences, 960 words, and 5919 characters; of different news articles picked from different newspapers like “The Hindu”, “Indian Express” etc. over different areas. The results of this search are shown in Table 1. The proposed system is showing the exact results of matching as compared with other systems. The computer assisted Proposed system with other available online tools is used to test the same content. Set Q consisted of 52 sentences, 923 words, and 6146 characters; where the text is picked from different topics on CS Subjects related to Engineering in terms of short assignment. Set R was the combination of 46 sentences, 992 words, and 6925 characters; from different Research papers from CS field. Set S is considered as a Partial set where sentences are picked from above three sets P, Q and T. Set S consisted of 30 sentences, 456 words and 2881 characters. Set T has “no plagiarism” throughout the detection and all three systems along with proposed system testing the same content. The results are purely accepted. Set T is prepared by own using 50 sentences, 626 words, and 3803 characters. Set U consisted of 51 sentences, 922 words, and 6114 characters. The content in this set is picked from the published paper on different journals related to field NLP, Networking, software Engineering, etc. The plagiarized percentage is shown in the line graph (figure 3). The last Set V consisted of 51 sentences, 837 words, and 5457 characters; where an equal proportion of text is picked from all the above sets which are a combination of news, computer science field assignment content, self-written content, research and published paper content. The proposed system is showing % of plagiarism towards the contents of different Sets. Line Graph for accuracy based comparison between different tools is shown below in terms of percentage:

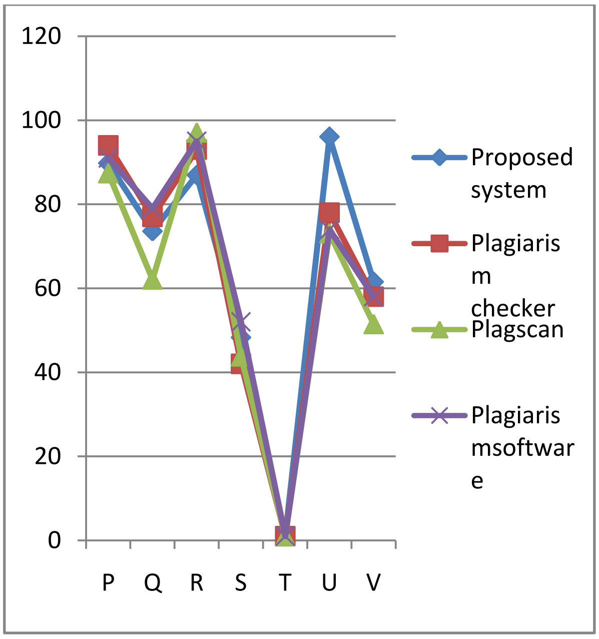

Figure 3: Accuracy comparison line graph

5. CONCLUSION

In the recent technology world, PD is crucial to sustaining the originality. In this document, this paper discussed types of plagiarism, along with text-based PD technique and its suitable approaches in addition to their level of use. The purpose of this document is to propose a system which is used for checking the plagiarized text by using the web (Google Search). Each individual sentence is searched on the internet by using the Cosine Similarity Approach. The output of this system is showing the plagiarized sentences with their original source (URL). The restriction of this system is a performance evaluation which is depending upon the speed of the internet and accuracy of the result by Google Search. In future work, the proposed system can be extended by using usability features that accuracy of this system can be improved. Plagiarism options can be extended with uploading of documents into different document type checking. The current system has only Google as a browsing search that can be extended to more options for the betterment of accurate results.

6. REFERENCES

[1] EncyclopediaWikipedia,https://en.wikipedia.org/wiki/Natural_language_processing (last accessed August 10, 2016).

[2] EncyclopediaMicrosoft,https://www.microsoft.com/en-us/research/group/natural-language-processing(Last accessed August 12, 2016)

[3] Encyclopedia,mind.ilstu,http://www.mind.ilstu.edu/curriculum/protothinker/natural_language_processing.php (Last accessed August 15, 2016).

[4] EncyclopediaBritannica,http://www.britannica.com/EBchecked/topic/462640/plagiarism (Last accessed August 18, 2016).

[5] Mechti, S., Jaoua, M. B., & Belguith, L. H. (2013). L H.: A framework for Plagiarism Detection based on Author Profiling.*Notebook for PAN at CLEF*.

[6] Joshi, M., & Khanna, K. (2013). Plagiarism detection over the web: review. *International Journal of Computer Applications*, *68*(15).

[7] Zechner, M., Muhr, M., Kern, R., & Granitzer, M. (2009, September).External and intrinsic plagiarism detection using vector space models. In*Proc. SEPLN* (Vol. 32, pp. 47-55.

[8] Zhang, P. Y., & Li, C. H. (2009, August).Automatic text summarization based on sentences clustering and extraction.In *Computer Science and Information Technology, 2009.ICCSIT 2009.2nd IEEE International Conference on* (pp. 167-170).IEEE.

[9] Clough, P. (2003).Old and new challenges in automatic plagiarism detection.In *National Plagiarism Advisory Service, 2003; http://ir.shef.ac. uk/cloughie/index.html*.

[10] Shivakumar, N., & Garcia-Molina, H. (1996, April).Building a scalable and accurate copy detection mechanism.In *Proceedings of the first ACM international conference on Digital libraries* (pp. 160-168).ACM.

[11] Ali, A. M. E. T., Abdulla, H. M. D., & Snasel, V. (2011).Overview and Comparison of Plagiarism Detection Tools. In *DATESO* (pp. 161-172).

[12] Si, A., Leong, H. V., & Lau, R. W. (1997, April).Check: a document plagiarism detection system. In *Proceedings of the 1997 ACM symposium on Applied computing* (pp. 70-77). ACM.

[13] Tripathi, R., Tiwari, P., & Nithyanandam, K. (2015, January).Avoiding plagiarism in research through free online plagiarism tools.In *Emerging Trends and Technologies in Libraries and Information Services (ETTLIS), 2015 4th International Symposium on* (pp. 275-280).IEEE.

[14] Monostori, K., Zaslavsky, A. B., & Schmidt, H. W. (2000, May). MatchDetectReveal: finding overlapping and similar digital documents. In*IRMA Conference* (pp. 955-957).

[15] HaCohen-Kerner, Y., Tayeb, A., & Ben-Dror, N. (2010, August).Detection of simple plagiarism in computer science papers.In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 421-429).Association for Computational Linguistics.