# SEMANTICALLY PLAGIARISM DETECTION SYSTEM USING WEB SERVICES

**Kamalpreet Sharma[1], Dr. Balkrishan Jindal[2]**

*[1]Student of Yadvindra College of Engineering Punjabi University Patiala,*
*Guru Kashi Campus, Talwandi Sabo, Punjab, (India)*

*[2]Assistant Professer of Yadvindra College of Engineering Punjabi University Patiala,*
*Guru Kashi Campus,Talwandi Sabo, Punjab, (India)*

## ABSTRACT

*Plagiarism is the "wrongful appropriation" and "stealing and publication" of another author's "language, thoughts, ideas, or expressions" and the representation of them as one's own original work. Plagiarism can also be hidden when text is translated from one language to another with no credit to the version, which is called cross- language plagiarism. Plagiarism is widely found in text, document's, papers, codes, images. Plagiarism detection systems find copies, not plagiarism and for translations or heavily edited material, the systems are powerless. We have proposed a new system by using web services for detecting different types of plagiarism, this system able to detect synonyms, translations of words.*

***Keywords: Intelligent Plagiarism, Literal Plagiarism, Semantic Analysis, Web Services***

## I. INTRODUCTION

Plagiarism can be defined as turning of someone else's work as our own without reference to original source. Plagiarism is copying another person's ideas, words or writing and pretending that they are one's own work. It can involve violating copyright laws. College students who are caught plagiarizing can be expelled from school, and writers who plagiarize will often be taken less seriously. Commonly in practice there are different plagiarisms methods like Copy – paste plagiarism (copying word to word textual information), Paraphrasing (copying same content in different words),Translated plagiarism (content translation and use without reference to original work), Artistic plagiarism (presenting same work using different media: text, images etc.), Idea plagiarism (using similar ideas which are not common knowledge), Code plagiarism (using program codes without permission or reference),No proper use of quotation marks (failing to identify exact parts of borrowed content),Misinformation of references (adding reference to incorrect or non existing source).

## 1.1 Types of Plagiarism

### 1.1.1 Literal Plagiarism

Literal plagiarism is a common and major practice wherein plagiarists do not spend much time in hiding the academic crime they committed. For example, they simply copy and paste the text from the Internet. Aside from few alterations in the original text, copy text word for word.

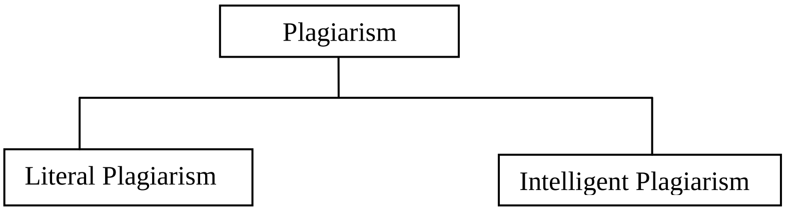


**Fig. 1 Types of Plagiarism**

### 1.1.2 Intelligent Plagiarism

Intelligent plagiarism is a serious academic dishonesty wherein plagiarists try to deceive readers by changing the contributions of others to appear as their own. Intelligent plagiarists try to hide, obfuscate, and change the original work in various intelligent ways, including text manipulation, translation and idea adoption.

**1.1.2.1 Text Manipulation:** Plagiarism can be obfuscated by manipulating the text and changing most of its appearance

**1.1.2.2 Translation:** Obfuscation can also be done by translating the text from one language to another without proper referencing to the original source.

**1.1.2.3 Idea Adoption:** Idea adoption is the most serious plagiarism that refers to the use of other's ideas, such as results, contributions, findings, and conclusions, without citing the original source of ideas.

## 1.2 Plagiarism Detection Methods

### 1.2.1 Fingerprinting

Fingerprinting is currently the most widely applied approach to plagiarism detection. This method forms representative digests of documents by selecting a set of multiple substrings (n-grams) from them.

### 1.2.2 Citation analysis

Citation-based plagiarism detection relies on citation analysis, and is the only approach to plagiarism detection that does not rely on the textual similarity. It examines the citation and reference information in texts to identify similar patterns in the citation sequences. As such, this approach is suitable for scientific texts, or other academic documents that contain citations.

### 1.2.3 Stylometry

Stylometry subsumes statistical methods for quantifying an author's unique writing style and is mainly used for authorship attribution. By constructing and comparing Stylometry models for different text segments, passages that are stylistically different from others, hence potentially plagiarized, can be detected.

### 1.2.4 String Matching

String matching is a prevalent approach used in computer science, compare original string with suspicious document to detect plagiarized string.

### 1.2.5 Bag of words

Bag of words analysis represent the adoption of vector space retrieval, a traditional IR concept, to the domain of plagiarism detection. Documents are represented as one or multiple vectors, e.g. for different document parts, which are used for pair wise similarity computations. Similarity computation may then rely on the traditional cosine similarity measure, or on more sophisticated similarity measures.

### 1.3 Parameters that Effect Plagiarism

**1.3.1 Synonyms:** A synonym is a word or phrase that means exactly or nearly the same as another word or phrase in the same language. For example, if we talk about a long time or an extended time, long and extended are synonymous with in that context.

**1.3.2 Active Passive:** A sentence whose agent is marked as grammatical subject is called an active sentence. In contrast, a sentence in which the subject has the role of patient or theme is named a passive sentence, and its verb is expressed in passive voice. E.g. (Harry ate six shrimp at dinner (active)) (At dinner, six shrimp were eaten by Harry (passive)).

**1.3.3 Tense:** In grammar, tense is a category that expresses time reference. Basic tenses found in many languages include the past, present and future, E.g. (The sun sets in the west) (All the cars stop at this crossing).

## II. RELATED WORK

Protection of digital documents from illegal copy has received much attention recently. Most of techniques for copy case detection are based on ideas of substring matching. In paper [3] string matching approach basically identifies maximum matches in pairs of strings, which not possible for large amount of data. In paper [4] and paper [5] keyword similarity mechanism is used but fail give any method for semantic plagiarism detection.

## III. METHODOLOGY

The user uploads the document to check the plagiarism in its document. The proposed system decomposed the document into single words. By web crawling process using web services, system performs the semantic comparison and also compares the user document with web to check plagiarism in the user document. After the comparison the user document result will be send to the user with highlighted link, it define source of data i.e. from where user document is copied. If no links are displayed then there is no plagiarism in user document.

3.1 Algorithm

**1.** User uploads documents in proposed software to detect plagiarism.

**2.** On client machine decomposition of documents is performed.

**3.** In next phase software check the document semantically.

**4.** Plagiarism Checker System performs searching using web services follows web crawling process.

**5.** Proposed System display results, if data is copied from any source gives hyperlinks and if not found then content is unique.

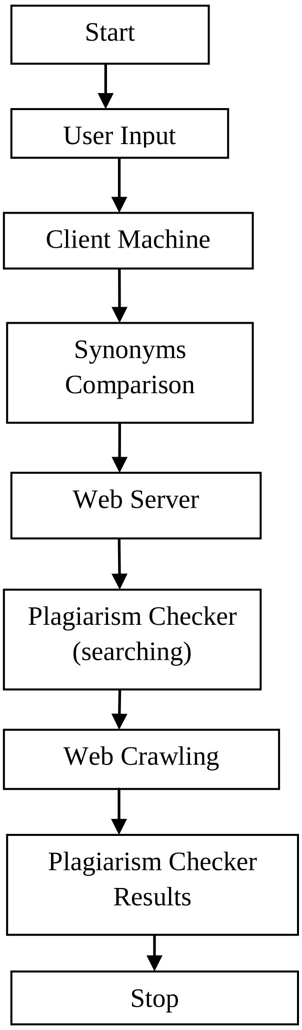


**Fig. 2 Shows Algorithm of Proposed Method**

## IV. RESULTS

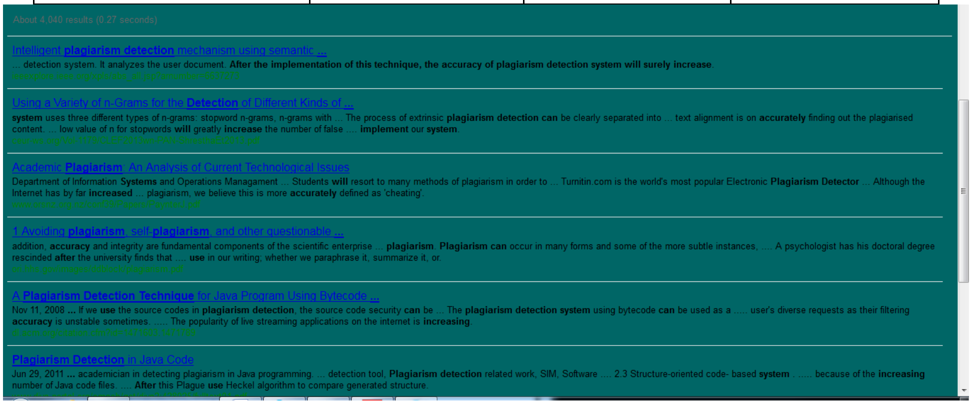
Proposed System display plagiarism detection search results shown in figure 3, it is based on multi gram approach. It check plagiarism through web, no local database is created. It split the document into single word. Proposed system gives the links of source from where data is copied. Through these formulas proposed system efficiency is calculated.

Sensitivity = No of links detected about plagiarized data/ Total no of Inputs

Specifity = No of links not detected about plagiarized data/ Total no of Inputs

**Table 1 Shows the Output of System**

| Category of Input Text | Sensitivity | Specifity | Efficiency |
|---|---|---|---|
| Thesis Files | 0.995 | 0.005 | 99.5 |
| College Assignments | 0.99 | 0.01 | 99.5 |
| Others | 0.990740741 | 0.009259259 | 99.07407407 |



About 4,040 results (0.27 seconds)

Intelligent **plagiarism detection** mechanism using semantic **...**
... detection system. It analyzes the user document. **After the implementation of this technique, the accuracy of plagiarism detection system will surely increase**.
ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6637273

Using a Variety of n-Grams for the **Detection** of Different Kinds of **...**
**system** uses three different types of n-grams: stopword n-grams, n-grams with ... The process of extrinsic **plagiarism detection can** be clearly separated into ... text alignment is on **accurately** finding out the plagiarised content. ... low value of n for stopwords **will** greatly **increase** the number of false .... **implement** our **system**.
ceur-ws.org/Vol-1179/CLEF2013wn-PAN-ShresthaEt2013.pdf

Academic **Plagiarism**: An Analysis of Current Technological Issues
Department of Information **Systems** and Operations Managament ... Students **will** resort to many methods of plagiarism in order to ... Turnitin.com is the world's most popular Electronic **Plagiarism Detector** ... Although the Internet has by far **increased** ... plagiarism, we believe this is more **accurately** defined as 'cheating'.
www.orsnz.org.nz/conf39/Papers/PaynterJ.pdf

1 Avoiding **plagiarism**, self-**plagiarism**, and other questionable **...**
addition, **accuracy** and integrity are fundamental components of the scientific enterprise ... **plagiarism**. **Plagiarism can** occur in many forms and some of the more subtle instances, .... A psychologist has his doctoral degree rescinded **after** the university finds that .... **use** in our writing; whether we paraphrase it, summarize it, or.
ori.hhs.gov/images/ddblock/plagiarism.pdf

A **Plagiarism Detection Technique** for Java Program Using Bytecode **...**
Nov 11, 2008 **...** If we **use** the source codes in **plagiarism detection**, the source code security **can** be ... The **plagiarism detection system** using bytecode **can** be used as a ..... user's diverse requests as their filtering **accuracy** is unstable sometimes. ..... The popularity of live streaming applications on the internet is **increasing**.
dl.acm.org/citation.cfm?id=1471603.1471789

**Plagiarism Detection** in Java Code
Jun 29, 2011 **...** academician in detecting plagiarism in Java programming. ... detection tool, **Plagiarism detection** related work, SIM, Software .... 2.3 Structure-oriented code- based **system** . ..... because of the **increasing** number of Java code files. .... **After** this Plague **use** Heckel algorithm to compare generated structure.

**Fig. 3 Showing Links of plagiarized data by proposed system for document**

The proposed method is compared with 100 thesis files, 105 college assignments and other source of data to detect plagiarism. The efficiency of proposed method is shown in figure 4, 99% efficient results as compare to existing system. It is a fast, accurate and efficient method for semantic plagiarism detection using web services.

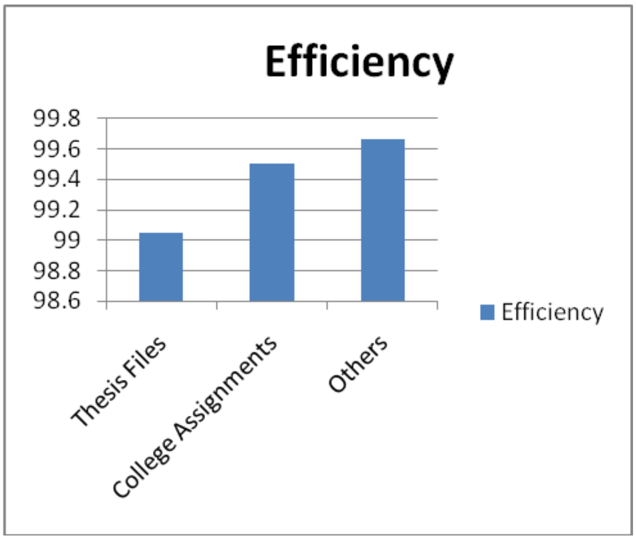


**Fig. 4 Shows the Efficiency of Proposed Method**

## V. CONCLUSION

The proposed system works on semantic plagiarism detection using web services, detect plagiarized data fast and accurately as compare to existing system. It removes the limitation of existing system as they are based on string matching and do not check semantic of words. It works on semantic technology.

## REFERENCES

[1] M.S. Alzahrani, N. Salim and V. Palade, Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model, 3, 2015, 248-268.

[2] N. More, and A. Bhootra, Plagiarism Detection in Source Code", International Journal for Innovative Research in Science and Technology, 1,2015, 109-112.

[3] J.D. Velasquez, and E. Marrese, Tools for external plagiarism detection in DOCODE, Proc. IEEE International Joint Conference on Web Intelligence and Intelligent Agent Technologies, Warsaw, 2, 2014, 296-303.

[4] N. Meuschke and B. Gipp, Reducing Computational Effort for Plagiarism Detection by using Citation Characteristics to Limit Retrieval Space, IEEE Joint Conference on Digital Libraries, London,8-12 Sept. 2014,197-200.

[5] S. Arrish, F.N Afif, A. Maidorawa and M. Salim, Shape-Based Plagiarism Detection for Flowchart Figures in Texts, International Journal of Computer Science and Information Technology, 6(1),2014, 113-124.

[6] S. Mulcahy and C. Goodacre, Opening Pandora’s Box of academic integrity: Using plagiarism detection software, International Journal of Computer Science and Information Technology, 6(1), 2014, 688-696.

[7] D. Ceglarek, Evaluation of The SHAPD2 Algorithm Efficiency in Plagiarism Detection, IEEE International Conference on Technological Advances In Electrical, Electronics and Computer Engineering, Konya, 9-11 May 2013, 465-470..

[8] J. Agarwal, R.H. Goudar, P. Kumar, K. Sharma, V. Parshav, R.Sharma, A. Srivastava and R. Rao, Intelligent Plagiarism Detection Mechanism using Semantic technology: A Different Approach, IEEE International Conference on Advances in Computing , Communication and Informatic, Mysore,22-25 Aug. 2013, 779-783.

[9] N. Idika, H. Phan and V. Mayank, Achieving Linguistic Provenance via Plagiarism Detection, IEEE International Conference on Document Analysis and Document, Washington, 25-28 Aug. 2013,648-652.

[10] Cosma and M. Joy, An Approach to Source-Code Plagiarism Detection and Investigation Using Latent Semantic Analysis, IEEE, 61(3), 2012, 379-394.