Plagiarism Detection in Web based Learning Management Systems and Intellectual Property Rights in the Academic Environment

Dinesh Kumar Saini
Faculty of Computing and IT, Sohar University, Oman
Faculty of Engineering and IT, University of Queensland, Australia

Lakshmi Sunil Prakash
Faculty of Computing and IT, Sohar University, Oman

ABSTRACT
This paper presents a background of a project undertaken at Sohar University to implement a web based plagiarism detection system for academic activities. Digital content in academic environments are created by academics and students. This Digital content is research publications and assessment submissions – assignments. Academic Digital content once created needs to be verified for non-violation of IPR issues and compliance to academic plagiarism policies followed by the institution.

General Terms
Computer Science, Software System, Web, Learning Management Systems, Text Mining, IPR.

Key Words
Plagiarism Detection, IPR, Web Based Learning Management Systems, Learning Digital Content Authentication, Digital Content

1. INTRODUCTION
Web based Learning Management systems are now a mandatory resource in any reputable academic institution. Institutions benefit a great deal from the flexibility that these systems provide to the academic and learner community. Learning Digital content can be made available 24x7 to any learner connected to the Internet from anywhere in the world. Academics can receive assessment submissions from off-campus and part-time students at their convenience. However these systems also present severe challenges to academics with regards to the following.

Verifying the integrity of learning - Digital content available in these systems - Some academics and Digital content creators unwittingly or knowingly use Digital content created by another person without properly citing them and passing it off as their own. Internet Digital content is easy to copy and therefore it is assimilated into a learning material without checking for integrity or authorship.[1,2]; mention cases of academics copying from conference and journal papers.

Verifying the quality of submissions received by academic is one of the toughest tasks in todays learning environment. Numerous researchers have documented the extent of plagiarism and student cheating over the past 60 years [2,3]. [4] observes that it is so easy to plagiarize using Internet sources, students may plagiarize without recognizing that they are doing so, knowing that plagiarism is ethically wrong. Both a) and b) come under the category of IPR (Intellectual Property Rights) regulations and plagiarism. An institution needs to adhere to the quality audit process that details the auditing mechanism that must be applied by the academics at the institution while dealing with Digital content in e-learning systems and student submissions.

Many educators acknowledge that more students than ever plagiarize material from different sources, especially the Internet [2]. [3, 5] observed that the information technology boom has attributed to the widespread practice of plagiarism.

The range of different types of plagiarism has been by mentioned in [6] namely copy and paste plagiarism, word switch plagiarism, but they essentially mean borrowing ideas without crediting the real author. Academic Institutions try to counter plagiarists by establishing strict academic integrity and anti-plagiarism policies [5, 7].One approach to fighting would be to change the campus culture from focusing plagiarism on "catching cheaters" to promoting academic integrity [8, 9].

2. FRAMEWORK OF PLAGIARISM DETECTION PLATFORM
The framework illustrated in Figure 1 is an illustration of the inputs to Digital content creation. The Digital content in academic environments is composed from the existing knowledge. This existing knowledge as to referenced/cited correctly so that new knowledge and existing knowledge have be appropriately differentiated. Citations and references also increase the academic value of Digital content.

The Digital content is then updated with new knowledge by the Digital content creator. New knowledge has to be ultimately verified, authenticated and validated by the repeat experiment, Hypothesis and Mathematical Induction for appropriate inclusion in the Digital content Management Systems.We are developing a framework that will help the academic community from content creation to the content dissemination.

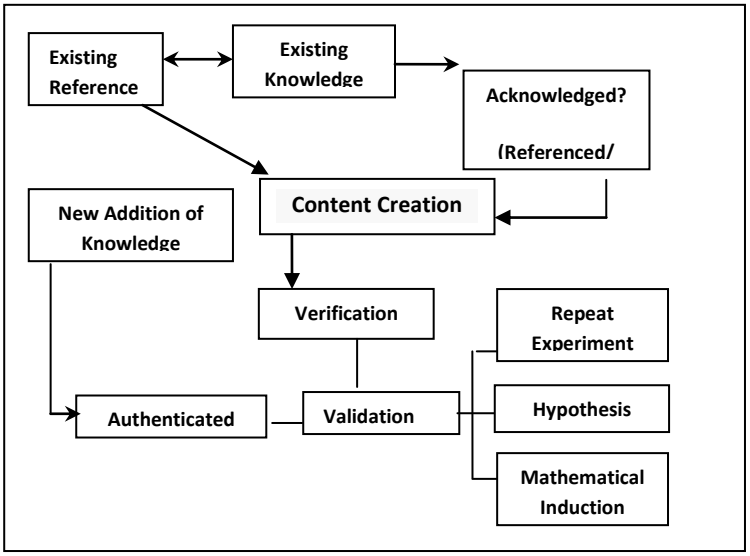

Fig. 1 Framework Design for Plagiarism detection Software

It is effortless to compose Digital content but to ascertain its knowledge value and its originality; the following steps need to be performed.

Comparison with similar Digital content in the domain.
Checking references/citations
Ascertaining existing knowledge
Highlighting the new knowledge /creativity/innovation in the composed digital content by matching existing digital content in the same domain.
Validation by Repeat Experiment, hypothesis and Mathematical Induction

3. DIGITAL CONTENT IN ACADEMIC ENVIRONMENTS

The Digital content in Web Based Learning Management Systems can be grouped under the following categories

Learning Digital content – Academics upload relevant course Digital content to students to complete their learning and this Digital content can be gleaned from existing knowledge available in websites, text books and other multimedia material. Referenced Digital content gives authenticity value to the Digital content. Such material will be reused and referenced several times.
Learning Digital content needs to be structured as

- Learning Objects- Context and Knowledge
- Student assessment submissions – archived and submitted

Web based learning management systems have content arranged as – learning objects. Learning objects have the following attributes Metadata, Rights, Technical, Educational, Relation, Annotation, General, Classification, Life Cycle.

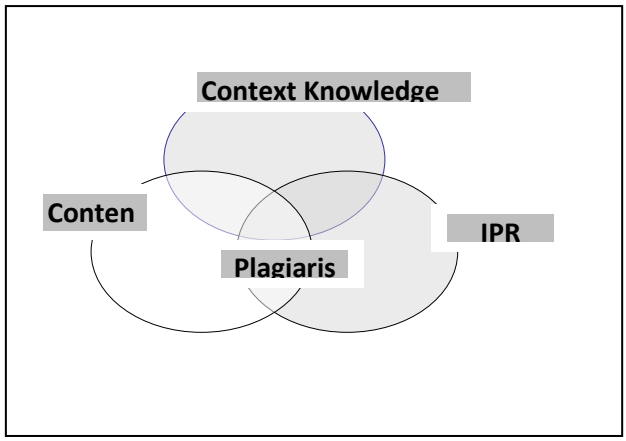

Fig. 2 Content, IPR, Context Knowledge and Plagiarism

Turnitin is good at Web searching almost all forms of Web documents except printed books and these. It's very presence in an academic institution deters students from plagiarism. It cannot interpret a citation instead it will just show the citation.

This paper presents a systematic framework using multilevel matching approach for plagiarism detection (PD). A multilevel structure, i.e. document–paragraph–sentence, is used to represent each document. In document and paragraph level, we use traditional dimensionality reduction technique to project high dimensional histograms into latent semantic space [16, 21].

4. ALGORITHMS USED IN PLAGIARISM DETECTION

The following algorithms have been used for plagiarism detection. The algorithms normally used in plagiarism detection software are string tiling, Karp-Rabin algorithm, Haeckel's algorithm, k-grams, string matching algorithm [10]. In [11], the authors describe two algorithms that they have used to test for efficiency in plagiarism detection.

In [12], the authors propose a system that is based on properties of assignments that course instructors use to judge the similarity of two submissions instead of the popular text-based analyses. This system uses neural network techniques to create a feature-based plagiarism detector and to measure the relevance of each feature in the assessment [13]. The system was trained and tested on assignments from an introductory computer science course, and produced results that are comparable to the most popular plagiarism detectors.

Two popular methods by Levenshtein [33] and Damerau [34] defined edit distances that can be used to compare the similarity of two strings of characters with each other. These distances are used in a variety of applications ranging from DNA analysis to plagiarism detection [35],[36].

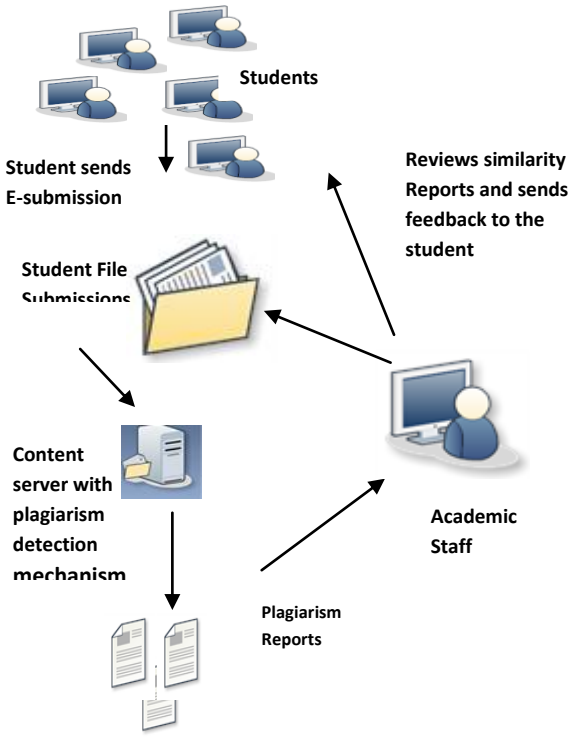

Fig. 3 Academic Plagiarism Scenario

The following are some of the work already done in this field. Scherbinin and Butakov [29] used *Levenshtein* distance to compare word *n*-gram and combine adjacent similar grams into sections. In another approach [30] the *Levenshtein* distance and simplified *Smith-Waterman* algorithm were merged as a single algorithm for the identification and quantification of local similarities in plagiarism detection. In [6] the researchers used the *LCS* distance combined with other POS syntactical features to identify similar strings locally and rank documents globally

A commonly-used bottom-up dynamic programming algorithm for computing the Levenshtein distance involves the use of an (n + 1) × (m + 1) matrix, where n and m are the lengths of the two strings. This algorithm is based on the Wagner-Fischer [32] algorithm for edit distance.

The following is an excerpt pseudocode for a function Levenshtein distance that takes two strings, s of length m, and t of length n, and computes the Levenshtein distance between them:

```
int LevenshteinDistance (char s[1…m], char
t[1…n])
// d is a table with m+1 rows and n+1
columns declare int d[0…m, 0…n] for i from
0 to m
  d[i, 0] := i
for j from 0 to n
   d[0, j] := j
for i from 1 to m
  for j from 1 to n
   if s[i] = t[j]
       then cost := 0                  else
       cost := 1
        d[I, j] := minimum(
       d[i-1,    j]    +1,
       d[i, j-1] +1,
       d[i-1, j-1] +cost
RETURN D[M, N]
```

The second algorithm, the Smith-Waterman algorithm [3] is a classical method of comparing two strings with a view to identifying highly similar sections within them. It is widely-used in finding good near-matches, or so-called local alignments, within biological sequences. But now, Smith-Waterman algorithm has been used in the text plagiarism detection, and our paper will simplify it [14, 15].

5. PLAGIARISM SOFTWARES

Software and websites both paid-for and free are available for use. However academically popular software are Turnitin and iThenticate. Turnitin was designed by John Barrie and a group of his UC Berkeley colleagues. This software uses digital technology to conduct a meta-search of the internet to locate sources from where the document might have been plagiarised. Teachers receive an Originality report that cites the degree of originality and links to Internet webpages that help determine what Web resources their students have tapped [17, 18].

Plagiarism.org was the largest plagiarism detection web site ten years ago. It was able to detect more complicated forms of plagiarism such as synonym replacement. However, its algorithm is still undisclosed, because it is the same one utilized today by plagiarism detection giant Turnitin.Com [10]. Plagiarism.org still exists today as a resource for all things plagiarism-related, but it directs its users to Turnitin.com for actual detection services. What is known about Plagiarism.org's, or now Turnitin.com's, design is that it has to maintain a mega-gigabyte database of documents. The results are given as a plagiarism index of zero to one; zero meaning no similarity, and one being a perfect copy [19, 20].

Most importantly, it differs from WordCheck and CopyCatch in that it is not based on counting occurrences of words. [10].

6. TEXT MINING

Text mining is a relatively new branch of data mining. In document comparison which is the focus of this paper, frequency modeling is a central aspect. Using frequency matrix, it is possible to prepare the matrix [22, 23].

In this paper, we introduce a duplicate detection scheme that is able to determine, with a particularly high accuracy, the degree to which one document is similar to another. Our pairwise document comparison scheme detects the resemblance between the content of documents by considering document chunks, representing contexts of words selected from the text. The resulting duplicate detection technique presents a good level of security in the protection of intellectual property while improving the availability of the data stored in the digital library and the correctness of the search results[13]. The key measures to calculate the measurement of resemblance and measure of content are discussed in this work. To do a comparison the documents are prepared for logical representation using chunking.

Textual data is a source of a large body of information. This information can be extracted from this data through various

Text mining offers a variety of approaches for extracting information and knowledge from textual data. One of the approaches discussed in [14] is that the attribution of authorship based on lexical and syntactical characteristics of the text can be used for detection of plagiarism and therefore has implications to the management of intellectual property rights.

The similarity between any two documents is the main concern of this investigation.

A solution to the outlined problems requires a reliable recognition of near-duplicates – preferably at a high runtime performance. These objectives compete with each other; a compromise in recognition quality entails deficiencies with respect to retrieval precision and retrieval recall. A reliable approach to identify two documents d and d as near-duplicates is to represent them under the vector space model, referred to as d and d , and to measure their similarity under the l –norm the enclosed angle

Near-duplicate detection is the task of identifying documents with almost identical content. The respective algorithms are based on fingerprinting; they have attracted considerable attention due to their practical significance for Web retrieval systems, plagiarism analysis, corporate storage maintenance, or social collaboration and interaction in the World Wide Web.

7. EXPERIMENT SETUP

Plagiarism is a global issue in academic environments. However the extent or effects of plagiarism has not been quantified. The aim of the research is to investigate the extent of plagiarism and to quantify the contents of plagiarism detection in web based learning environment using statistical analysis. The effectiveness of the usage of the plagiarism detection tools will also be analysed in the close domain of computing students. The extent of use of internet resources will be investigated an Students need to understand and appreciate the importance of submitting original work as part of their academic submissions. Currently, the learning management system is the main vehicle for academic activity in most reputed academic institutions. The study will review the plagiarism level in student submissions before and after integrating the anti-plagiarism tool.

By using plagiarism checker / detection tools, academics can be freed from the task of manual correction of students' work. This will help an institution to direct its academic staff to have a monitoring mechanism for checking the quality of student's submission. At the same time raise the level of awareness among academics fraternity regarding university anti-plagiarism policies and guide them towards higher academic standards. Learners will feel gratified that original work will be distinguished and rewarded. Integrating anti-plagiarism detection software into a learning management system will strengthen the academic integrity of an institution and verify the quality of its academic functions.

8. SIGNIFICANCE OF RESEARCH

In academic institutions around the world Web based Learning Management Systems present the following challenges to academics with regards to the following.

Verifying the quality of submissions received from students- Numerous researchers have documented the extent of plagiarism and student cheating over the past 60 years, observes that it is so easy to plagiarize using Internet sources; students may plagiarize without recognizing that they are doing so despite knowing that plagiarism is ethically wrong. Many educators acknowledge that more students than ever plagiarize material from different sources, especially the Internet observed that the information technology boom has attributed to the widespread practice of plagiarism. The range of different types of plagiarism has been by mentioned in namely copy and paste plagiarism, word switch plagiarism, but they essentially mean borrowing ideas without crediting the real author. Academic Institutions try to counter plagiarists by establishing strict academic integrity and anti-plagiarism policies One approach to fighting plagiarism would be to change the campus culture from focusing on "catching cheaters" to promoting academic integrity [27].

The aim of this project study would be to verify this at SU Campus.

In the pilot study, a set of students from Level 2, Level 3 and 4 will be selected for analysing the extent of plagiarism – with respect to the least percentage of plagiarism to the highest extent of plagiarism in each level. The study will also ascertain the type of plagiarism benefits by quantification to the learner and academic communities at the university.

•Innovation
Though there are discussions on the effectiveness of plagiarism detection tools (Lukashenka 2007), this study will try to generate statistics to support this statement through surveys and analysis of quantitative data.
•Goals & Objectives of the experiment
Goals –The main goal is to quantify the effectiveness of the plagiarism detection tools. We will review the plagiarism extent in the FCIT for the academic submissions and the effect of this tool. Second goal is to improve the quality of graduate submissions and research publications at Sohar University.

Objectives -Increase awareness about IPR issues and the importance of plagiarism detection tools in today's academic environments. This investigation will help the academics in the university to devise strategies to prevent plagiarism. We will create awareness among the academic fraternity and student community on the effectiveness and benefits of using plagiarism detection tools in assessing course assessments especially written assignments through seminars.

9. INTELLECTUAL PROPERTY RIGHTS (IPR)

Intellectual property rights of a creator/ contributor must be protected in any domain. IPR will be protected with the use of the proposed tools integrated with the web based learning environments.

All the latest inventions, literary work, artistic work and other stuff like names, images and designs used commercially comes under the intellectual property which is divided into two parts. One is commercial and other is copyright [26].

All the patents, industrial issues of designs and other related work of commercial relations is divided into industrial property. The other type is copyright, which includes literary and artistic works such as novels, poems and plays, films, musical works, artistic works such as drawings, paintings, photographs and sculptures, and architectural designs''. These

definitions are given by World Intellectual Property Organization (WIPO).

The strategic benefits of the IPR are as following:
• Innovation:
It support the inventions and inventors and creative expressions, it helps in setting the common standards for the creativeness support.

IPR helps in building the professional networks and supports the corporate relationships, and the community. Through IPR, collaborative agreements (e.g., joint ventures, strategic alliances) can be formulated; IPR can help in building informal relationships with industry networks which in return give some service to the community.

• Market positioning: IPSs helps in increasing market share through building a broader user base or securing market protection schemes and professional recognition or brand recognition; competitive signalling in the market. It helps the researchers and firms for the growth of technology and methodologies.

• Finance: IPSs helps in generating direct income from market transactions to cover R&D or for profit; increasing ability to rise venture capital; and cost cutting in the research environment.

Research organisations use various types of IP nonexclusively, and their choice of protection is related to the strategic benefits they are seeking, the operational aspects of their knowledge creation processes, and the characteristics of the research organisations and sectors in question.

IPRs knowledge and related issues must be known to the inventors and researchers so that the cost and supports for the inventions can be covered from the income generated through the inventions and research work.

10. CONCLUSION

The plagiarism in the academic world is universal problem and we attempted to analyze it and integrate in the Web Based System of the university called SULMS (Customized Moodle). It is a detection system for academic activities. This Digital content is research publications and assessment submissions – assignments. Academic Digital content once created needs to be verified for non-violation of IPR issues and compliance to academic plagiarism policies followed by the institutions. This effort will help the academic community in providing healthy practice of research and study among the academicians and students.

11. ACKNOWLEDGEMENT

The authors of the paper are thankful for the University Research Council and SURGE grant of Sohar University, Oman. This research is the part of the SURGE grant provided by the Sohar University. Thanks to Prof Lance Bode and Prof Tony for the Valuable Input.

12. REFERENCES

[1] C. H. Huang, Chu, S. S., and Guan, C. T, " Implementation and performance evaluation of parameter improvement mechanisms for intelligent e-learning systems," *Computers & Education*, vol. 49, pp. 597 - 614, 2007.

[2] A. U. Szabo, J, "Cybercheats: Is Information and Communication Technology Fuelling Academic Dishonesty?," *Active Learning in Higher EducationEducation þ Training*, vol. 5, pp. 80-99, 2004.

[3] M. Hart, & Friesner, T, "Plagiarism and poor academic practice - A threat to the extension of e-learning in higher education?," *Electronic Journal of e-Learning*, vol. 2, 2004.

[4] J. Kraus, "Rethinking plagiarism: What our students are telling us when they cheat.," *Issues in Writing*, vol. 13, pp. 80-95, 2002.

[5] J. A. Liles, & Rozalski, M.E., "It's a Matter of Style: A Style Manual Workshop for Preventing Plagiarism," *College & Undergraduate Libraries*, vol. 11, pp. 91-101, 2004.

[6] C. Barnbaum, "Plagiarism: A Student's Guide to Recognizing It and Avoiding It," 2002.

[7] J. M. C. Hughes, & McCabe, D. L. , "Understanding academic misconduct.," *Canadian Journal of Higher Education*, vol. 36, pp. 49-63, 2006.

[8] M. Devlin, "Policy, preparation, and prevention: Proactive minimization of studentplagiarism.," *Journal of Higher Education Policy & Management,*, vol. 28, pp. 45-58, 2006.

[9] T. B. Gallant, & Drinan, P "Organizational theory and student cheating:Explanation, responses, and strategies.," *Journal of Higher Education,*, vol. 77, p. 839, 2006.

[10] P. Clough. (June2000, Plagiarism in natural and programming languages: an overview of current tools and technologies,.

[11] B.-R. A. Z. Su, K.-Y. Eom, M.-K. Kang, J.-P. Kim,and M.-K. Kim,, "Plagiarism detection using the levenshtein distance and smith-waterman algorithm," in *ICICIC '08 Proceedings of the 2008 3rd International Conference on Innovative Computing Information and Control.* , Washington, DC, USA, 2008, p. 569.

[12] V. L. S. Engels, and M. Craig. , "Plagiarism detection using feature-based neural networks.," in *Proceedings of the Thirty-Eighth SIGCSE Technical Symposium on Computer Science Education*, Covington, Kentucky, March 2007, pp. 34-38.

[13] R. M. Federica Mandreoli, Paolo Tiberio, "A document comparison scheme for secure duplicate detection," *Int J Digit Libr* vol. Springer-Verlag 2004, 2004.

[14] M. M. Edda Leopold, and Gerhard Paa, "DATA MINING AND TEXT MINING FOR SCIENCE & TECHNOLOGY RESEARCH " in *Handbook of Quantitative Science and Technology Research*, H.F. Moed et al. (eds.), Ed., ed Printed in the Netherlands: © Kluwer Academic Publishers., 2004, pp. 187-213. Barnbaum, C. (2002). "Plagiarism: A Student's Guide to Recognizing It and Avoiding It."

[15] Devlin, M. (2006). "Policy, preparation, and prevention: Proactive minimization of studentplagiarism." Journal of Higher Education Policy & Management, 28(1): 45-58.

[16] Dinesh Kumar Saini, L. S. P. a. W. M. O. (2010). Review of Technological Challenges in Web - Based Learning Content Management Systems (LCMS) with special emphasis on extraction of Learning Contents.

International Symposium, College of Applied Science, Ministry of Higher Education. Oman: 43-49.

[17] Gallant, T. B., & Drinan, P (2006). "Organizational theory and student cheating:Explanation, responses, and strategies." Journal of Higher Education, 77(5): 839.

[18] Hart, M., & Friesner, T (2004). "Plagiarism and poor academic practice - A threat to the extension of e-learning in higher education?" Electronic Journal of e-Learning 2(1).

[19] Hughes, J. M. C., & McCabe, D. L. (2006). "Understanding academic misconduct." Canadian Journal of Higher Education 36(1): 49-63.

[20] Kraus, J. (2002). "Rethinking plagiarism: What our students are telling us when they cheat." Issues in Writing 13(1): 80-95.

[21] Lakshmi Sunil Prakash, Dinesh Kumar Saini, N.S. Kutti (2009). " Integrating EduLearn learning content management system (LCMS) with cooperating learning object repositories (LORs) in a peer to peer (P2P) architectural framework." ACM SIGSOFT Software Engineering Notes 34(3): 1-7.

[22] Liles, J. A., & Rozalski, M.E. (2004). "It's a Matter of Style: A Style Manual Workshop for Preventing Plagiarism." College & Undergraduate Libraries 11(2): 91-101.

[23] Lukashenko, R., Graudina, V., Grundspenkis, J.: (2007). Computer Based Plagiarism Detection Methods and Tools: an Overview. CompSysTech '07:Proceedings of the 2007 International Conference on Computer Systems and Technologies,, New York, ACM Press.

[24] Mark Pedersen, D. E., Samir K. Amin and Lakshmi Prakash (2004). Parsing Arabic relative clauses: a Paninian dependency grammar approach. Multitopic Conference, 2004. Proceedings of INMIC 2004. 8th International Proceedings of the 8th IEEE International Multi-topic conference

[25] Mark Pedersen, D. E., Samir K. Amin and Lakshmi Prakash (2004). Relative Clauses in Hindi and Arabic: a Paninian Dependency Grammar Analysis. Proceedings of the Workshop on Recent Developments in Dependency Grammar -COLING2004 – The 20th International Conference on Computational Linguistics, Geneva,Switzerland.

[26] N.S Kutti, K., Z,Sunil. L (2007). EduLearn:An e-Learning Architecture for Prototyping Web-Based Learning Systems. EISWT, Florida,USA.

[27] Saini, D. K., Sunil Prakash, Lakshmi ,Omar M Wail (2010). Review of Technological Challenges in Web - Based Learning Content Management Systems (LCMS) with special emphasis on extraction of Learning Contents. Proceedings of First Joint scientific SYMPOSIUM of the Colleges of APPLIED SCIENCES

[28] Szabo, A. U., J (2004). "Cybercheats: Is Information and Communication Technology Fuelling Academic Dishonesty?" Active Learning in Higher Education Education þ Training 5(2): 80-99.

[29] V. Scherbinin and S. Butakov, "Using Microsoft SQL server platform for plagiarism detection," in Proc. SEPLN, Donostia, Spain, pp. 36–37.

[30] Z. Su, B. R. Ahn, K.Y. Eom, M. K. Kang, J. P. Kim, andM. K. Kim, "Plagiarism detection using the Levenshtein distance and Smith-Waterman algorithm," in Proc. 3rd Int. Conf. Innov. Comput. Inf. Control, Dalian, Liaoning, China, Jun. 2008, p. 569.

[31] M. Elhadi and A. Al-Tobi, "Duplicate detection in documents and webpages using improved longest common subsequence and documents syntactical structures," in Proc. 4th Int. Conf. Comput. Sci. Converg. Inf. Technol., Seoul, Korea, Nov. 2009, pp. 679–684.

[32] R.A. Wagner and M.J. Fischer , The string to string corection problem,J.Assoc.Comput.Mach.21,No.1 (1974),168-183.