International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 05 | May -2017 www.irjet.net p-ISSN: 2395-0072

# An Automatic Extraction of Educational Digital Objects and Metadata from institutional Websites

**Kajal K. Nandeshwar[1], Praful B. Sambhare[2]**

*[1]M.E. IInd year, Dept. of Computer Science, P. R. Pote College of Engg, Amravati, Maharashtra, India*
*[2]Assistant Professor, Dept. of Computer Science, P. R. Pote College of Engg, Amravati, Maharashtra, India*

---

***Abstract:*** *Among others things, as an educational information source, internet is used. In this project, a tool is provided which can detect all educational digital objects in any format that are already published on institutional websites and can be uploaded to a repository. This recopilation is a tedious task and is usually performed manually. In this project, the proposed system architecture is for automating this task of collecting the documents within any educational web domain and detects the documents that are loaded into a repository. In addition, its metadata such as abstract, author name is automatically extracted if available. The aim of the proposed system is automatically extracts the EDOs that are uploaded on the institutional websites and stored into the repository.*

***Keywords:*** **EDOs, automatic, website, Repository, links, extraction, information gathering**

## 1. INTRODUCTION

Internet is a powerful source of information which is used as an educational information source. Educational resource also called as digital object, learning object, learning resources, digital resources, digital content, reusable learning object, educational content in the field of technology enhanced learning (McGreal, 2004). Nowadays one of the most important sources of educational material is web where students and teachers have a large amount of information at their disposal [4]. An Educational Digital Object (EDO) is any material in the digital format that can be used as educational resource. For example, a scientific publication, an educational material that is used in a class is an educational resource [1]. Manual data extraction process is time consuming and error prone. Web pages come in the different formats including text, HTML pages, PDF documents, and other proprietary formats. Web pages may give the same or analogous information utilizing entirely diverse formats or linguistic uses, which makes addition of the information a fascinating task [14]. The web links provide a source of valuable information. In this project, system architecture is used for collecting the documents to assist the manager of institutional repositories in the recopilation task of EDOs within a website. Thus, plausible documents to be uploaded to a repository can be detected. Also, its metadata such as abstract, author name, affiliation if available are automatically extracted.

A problem that can be found in this extraction of EDOs is that many times, the required data are not in the document. These data can be in the different pages of the same website. The proposed system architecture takes advantage of this feature to improving the automation of information extraction. Therefore, in this system some data extracted are searched in the document and also searched in another page of the same sites. The proposed system gathers the EDOs and metadata which is in the form of list of links that are uploaded on the institutional website or any website and stored into the repository. The system receives as input URL of website or a text where the search is performed. The output of the system shows the retrieved documents together with the extracted information in a database.

## 2. EXISTING WORK

Automatic extraction plays an important role in processing results from search engines [7]. Regarding to the automatic gathering information systems, various proposals have been developed.

DeLa (Data Extraction and Label Assignment for Web Databases): DeLa describe by J. Wang and F.H. Lochovsky [9] which automatically extracts data from website and assigns the meaningful labels to the data. This technique concentrates on the pages that querying back end database using the complex search forms other than using keywords.

ViPER (Visual perception based Extraction of Records): It is described by K. Simon and G. Lausen [13] which is a totally automated information extraction tool. This technique is based on the assumption that the web page contains at least two consecutive data records which exhibits some kind of the structural and visible similarity. ViPER is able to extract the relevant data with respect to user's visual perception of the webpage. It only extracts the contiguous page in a website and it fails to perform nested structure effectively. It performs the good data extraction but implementation is not available [12].

AGATHE: The Agathe described by Albitar, Espinasse, and Fournier [2] that proposes generic multi agent architecture for the contextual information gathering on web domains. In Agathe, software agents exploit ontologies in order to realize web page classification and the task of information extraction. The AGATHE system is a generic software architecture allowing the development of the information gathering systems on the Web, for one or more restricted domains.

CROSSMARC: The CROSSMARC [5] is a project of multiple domain system which support the development of an agent based multilingual system for the extraction of information from the web pages. It uses an approach which is based on the knowledge combined with the machine learning techniques for designing a robust system in order to information extracting from interested websites. CROSSMARC reduces the high cost of maintaining the system. Because of the constant change of the web, this hybrid approach supports adaptability to new emerging concepts and a degree of independence from the specific web sites considered in the training phase [1].

NET (Nested data extraction using Tree matching and Visual cues): NET is proposed by Bing Liu and Yanhong Zhai [11] which extract data items from data records even it also handles nested data records. Building the tag tree is difficult because of page may contain erroneous and the unbalanced tags. This is performed based on nested rectangles. In order to find nested data records which are found at lower level, NET traverses the tag tree in post order (bottom up).

CiteSeerX: It is the next generation of the CiteSeer architecture which is a scientific literature digital library and search engine which automatically crawls and indexes the scientific documents in the computer science field. Its architecture based on the modular web services, pluggable service components, distributed object repositories and transaction safe processes. Its architecture enhances the flexibility, scalability and performance [6]. CiteSeer automatically retrieves and discovers online scientific documents. However, all these works consider documents that have some type of structure such as Call for Papers or scientific papers [1][3].

IEPAD: IEPAD described by C. H. Chang and S.C. Lui [10] which is an information extraction system which applying pattern discovery techniques. The IEPAD can automatically identify record boundary by repeated pattern mining and multiple sequence alignment. The discoveries of the repeated patterns are realized through a data structure call PAT trees. Additionally, repeated patterns are extended by the pattern alignment to comprehend all record instances.

## 3. PROPOSED SYSTEM

### 3.1 System Architecture of Proposed System

The architecture of the proposed system containing a query which takes text (it can be a URL of the website or combination of the text and website) as an input and then the input text is given to NLP and NLP (Natural Language Processing) technique is used for removing the un useful words from the text. With NLP, the text is broken into tokens (as in programming language parsing) at the lowest level. From there sentences can be identified. Within sentences user can look to determine likely context of words and phrases, using various dictionaries and domain specific lexicons. By this stage user should have recognized tokens that are proper names or other useful information. Then new text is provided to the crawler and then the crawler crawling website and extracting its contents. Usually websites are designed for visualization not for data exchange. A Web crawler is a program which inspects the web pages in a methodical and automated way [8]. Web crawling is the process used by search engines to collect pages from the Web. One of its uses is to create a copy of all visited web pages by a search engine that indexes pages providing a fast search for later processing. The beginning of the Web crawlers is visiting a list of URLs, identify the links in these pages and add them to the list of URLs to visit recurrently according to a given set of rules. The usual processing of a crawler is from a group of initial URLs addresses where linked resources are downloaded [1].

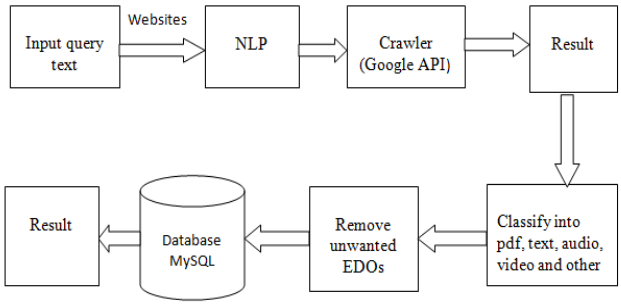



**Fig -1**: System Architecture of proposed system

Then the result which is in the form of links is produces by the crawler. Then EDOs are collects from the crawler and the metadata related to the EDOs are extracted. Then these are classified by the classifier which extracted contents into the audio, video, text, pdf, ppt, etc. Then the unwanted EDOs such as duplicate EDOs and other which are not useful are removed from the list. Then all EDOs and its metadata are saved to the database for later use. Then what user needs like audio, video, text, ppt, pdf, etc. and extract the data. The output of the system shows the list of EDOs in the form of links and

also shows how many links are fetched and the time needs to fetch those links.

### 3.2 Flowchart of Proposed System

The flowchart provides the basic flow of execution for the implementation of current system. The flowchart of the proposed system starts with entering a text and it can be a URL of the website or combination of the text and website and then the crawler crawling website and extracting its contents. Then EDOs are collects from the crawler and then these are classified by the classifier which extracted contents into the pdf, ppt, audio, video, text, etc. After that the unwanted EDOs are discarded from the list and all EDOs and metadata are saved to the database. Then what user needs like pdf, audio, video, and text and extract the appropriate data.

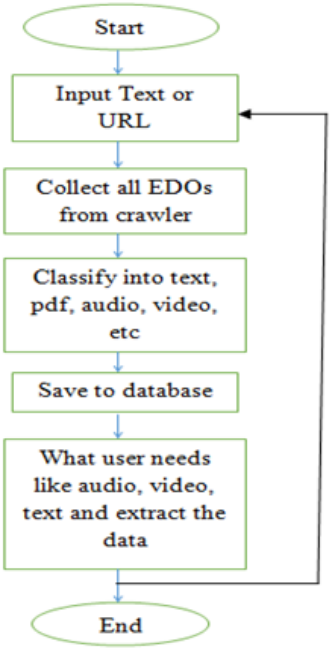



**Fig -2**: Flow chart of proposed system

## 4. RESULT ANALYSIS

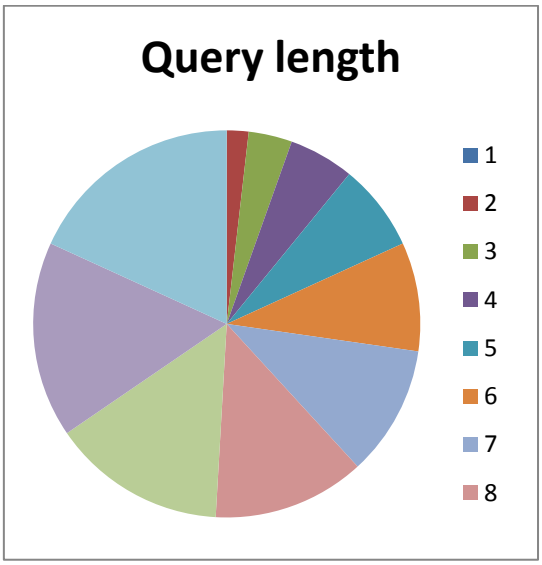
The proposed work builds in java framework uses MySQL for backend process. The table have been created for maintain the record of all EDOs of related search. This work is based on collecting the all EDOs and metadata (if available) which are uploaded on the institutional websites or on any web domain and stored into the repository. The all EDOs and metadata are extracted from the web domain and stored into the database and all result is in the form of links. The system shows the how many number of links fetched and time needed for fetching those links in ms.

Also when user searches the educational objects of the respective search second time or so on then the EDOs are extracted from the web domain as well as from the database where previous EDOs are stored and provides the output from both previous and new data. Using Natural Language Processing technique the time required for extracting the EDOs is minimized. The graph of extracting the EDOs with using NLP and without using NLP is shown for analyzing it. The experimental result of proposed work is shown below in the following graphs.

**Table -1:** Extracting EDOs with and without using NLP

| Query length | links | | delay(ms) | |
|---|---|---|---|---|
| | with NLP | without NLP | with NLP | without NLP |
| | | | | |
| 1 | 90 | 88 | 15955 | 17521 |
| 2 | 71 | 72 | 13427 | 15392 |
| 3 | 88 | 129 | 14049 | 20909 |
| 4 | 53 | 62 | 13227 | 13668 |
| 5 | 79 | 80 | 18574 | 21908 |
| 6 | 103 | 95 | 19359 | 23215 |
| 7 | 77 | 78 | 18898 | 22981 |
| 8 | 69 | 30 | 14259 | 17852 |
| 9 | 58 | 67 | 17651 | 21803 |
| 10 | 71 | 67 | 20278 | 22871 |

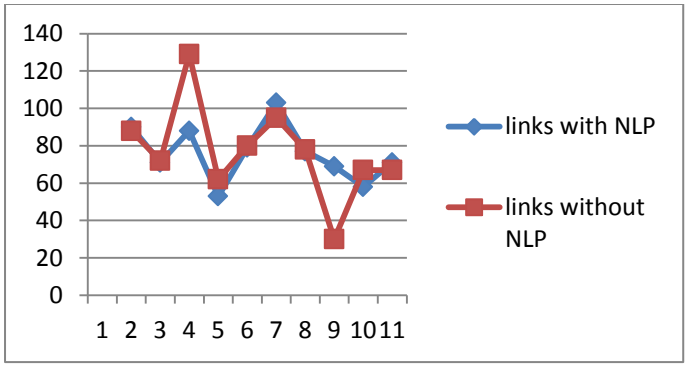
From the above result we can see that the query length 1 is containing the more number of links in less time with using the NLP than without using NLP. Also the query length 6, 8 and 10 are containing the more number of links in less time with using NLP. In the above table delay are shown in ms.




**Graph-1:** Graph of query length of search

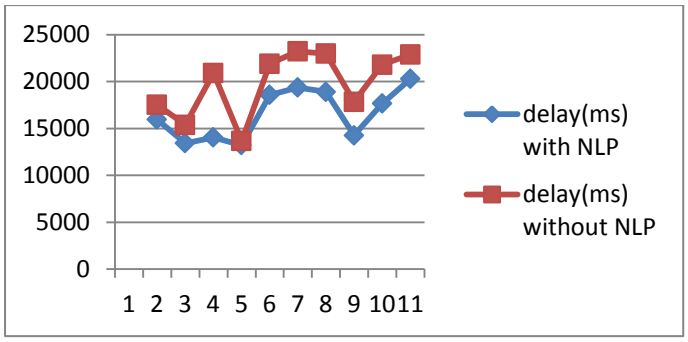
The above Pie chart indicates the various length of query of the search. The query length is depends on user that

how many words are entered as an input text. The query length from one to ten is shown in the above graph.




**Graph-2:** Comparison graph of links fetched with and without NLP

From the above graph the comparison factor between number of links fetch with using NLP and number of links fetch without using NLP is shown. On X-axis the query length is provided and on Y-axis the number of links fetched is provided.




**Graph-3:** Comparison graph of delay with and without NLP

From the above graph the comparison factor between time needed for fetching the links with using NLP and time needed for fetching the links without using NLP is shown. On X-axis the query length is provided and on Y-axis the delay time (ms) is provided.

**Table -2:** Table for Deep Learning Technique

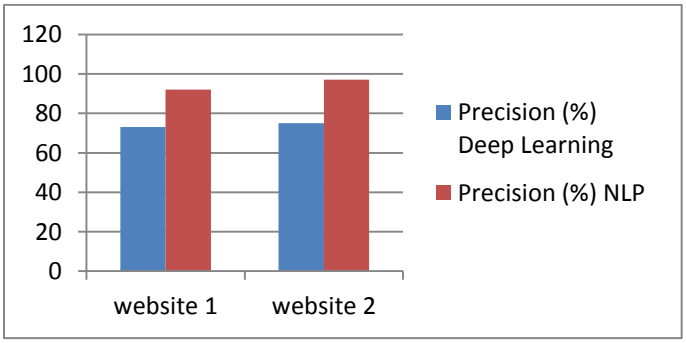
| Website | Precision (%) | Recall (%) |
|---|---|---|
| http://www.international.ucla.edu/korea/ | 73 | 75 |
| http://www.coronaregional.com/ | 75 | 76 |

The table contains the precision and recall value for the websites by Deep Learning technique. It contains the precision and recall value for first website and for the second website.

**Table-3:** Table for NLP Technique

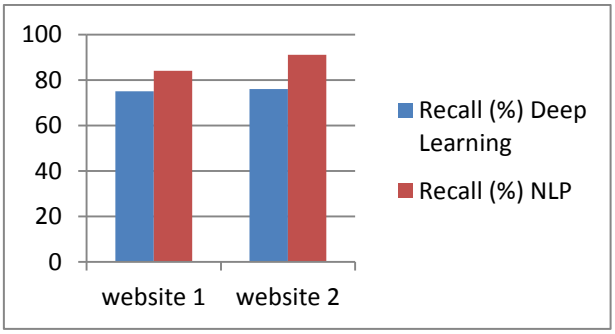
| Website | Precision (%) | Recall (%) |
|---|---|---|
| http://www.international.ucla.edu/korea/ | 92 | 84 |
| http://www.coronaregional.com/ | 97 | 91 |

The above table contains the precision and recall value for the websites by the NLP technique. It contains the precision and Recall value for first website and precision and recall values for the second website.




**Graph-4:** Comparison graph of precision by Deep Learning and NLP

The above graph contains the comparison factor on precision value between the Deep Learning technique and NLP technique. On the X-axis the websites are provided and on the Y-axis the precision values for website 1 by using Deep Learning and NLP technique and recall values for website 2 by using Deep Learning and NLP technique are provided. In the above graph, the NLP contains better precision values for website 1 and website 2.




**Graph-5:** Comparison graph of Recall by Deep learning and NLP

The above graph contains the comparison factor on recall value between the Deep Learning technique and

NLP technique. On the X-axis the websites are provided and on the Y-axis the recall values for website 1 by using Deep Learning and NLP technique and recall values for website 2 by using Deep Learning and NLP technique are provided. In the above graph, the NLP contains better recall values for website 1 and website 2.

## 5. CONCLUSIONS

The proposed system provides the collection of all EDOs and metadata when available in any formats that are uploaded on institutional websites and can be stored to a repository. The proposed system improves the automation of extraction of data. Therefore, in this system some data extracted are searched in the document text and are also searched in other pages of the same website. It provides the classified documents such as in pdf, word doc, text, ppt, video, etc and all documents in pdf format are placed in one place and so on so that user can easily found the particular format data. The proposed system provides the any documents that are on the any web domain at a single search.

## REFERENCES

[1] A. Casali, C. Deco and S. Beltramone, "An assistant to populate Repositories: Gathering Educational Digaital Objects and Metadata Extraction", in IEEE Revista Iberoamericana De Technologias Del Aprendizaje, vol.11, No.2, May 2016.

[2] S. Albitar, B. Espinasse, and S. Fournier, "Combining agents and wrapper induction for information gathering on restricted Web domains," in Proc. 4th Int. Conf. Res. Challenges Inf. Sci., Nice, France, May 2010, pp. 343–352.

[3] H. Li et al., "CiteSeer: A scalable autonomous scientific digital library," in Proc. 1st Int. Conf. Scalable Inf. Syst., Hong Kong, 2006, p. 18-es, doi: 10.1145/1146847.1146865.

[4] A. Casali, C. Deco, A. Romano, and G. Tomé, "An assistant for loading learning object metadata: An ontology based approach," Interdiscipl. J. E-Learn. Learn. Objects, vol. 9, pp. 77–87, Jan. 2013.

[5] M. T. Pazienza, A. Stellato, and M. Vindigni, "Combining ontological knowledge and wrapper induction techniques into an e-retail system," in Proc. Int. Workshop Tutorial Adapt. Text Extraction Mining (ATEM), Cavtat, Croatia, 2003, pp. 50–57.

[6] C. L. Giles, K. Bollacker, and S. Lawrence, "CiteSeer: An automatic citation indexing system," in Digital Libraries 98 - The Third ACM Conference on Digital Libraries, I. Witten, R. Akscyn, and F. M. Shipman III, Eds. Pittsburgh, PA: ACM Press, June 23–26 1998, pp. 89–98.

[7] Devika k, Subu Surendran, "An Overview of Web Data Extraction Techniques", International Journal of Scientific Engineering and Technology (ISSN : 2277-1581) Volume 2 Issue 4, 1 April 2013, pp: 278-287.

[8] C. Castillo, "Effective Web crawling," Ph.D. dissertation, Dept. Comput. Sci., Univ. Chile, Santiago, Chile, Nov. 2004.

[9] J. Wang and F.H. Lochovsky, "Data Extraction and Label Assignment for Web Databases", Proc. international conference on World Wide Web (WWW-12), pp. 187-196, 2003.

[10] C.H.Chang and S.C.Lui," IEPAD: Information Extraction Based on Pattern Discovery", Proc. International Conference on World Wide Web (WWW-10), pp. 223-231, 2001.

[11] Bing and Yanhong Zhai, "NET - A System for Extracting Web Data from Flat and Nested Data Records", Proc. WISE'05 Proceedings of the 6th international conference on Web Information Systems Engineering, pp. 487-495, 2005.

[12] Sureshkumar .T et.al, "A SURVEY OF TOOLS FOR EXTRACTING AND ALIGNING THE DATA IN WEB", International Journal of Computer Science & Engineering Technology (IJCSET) ISSN : 2229-3345 vol.5 No.03 Mar 2014, pp. 262-265.

[13] K. Simon and G. Lausen, "ViPER: Augmenting Automatic Information Extraction with Visual Perceptions",Proc. International Conference on Information and Knowledge Management (CIKM), 2005.

[14] J. Sharmila and a. subramani, "A COMPARATIVE ANALYSIS OF WEB INFORMATION EXTRACTION TECHNIQUES DEEP LEARNING vs. NAIVE BAYES vs. BACK PROPAGATION NEURAL NETWORKS IN WEB DOCUMENT EXTRACTION", ICTACT Journal on Soft Computing, Volume: 06, Issue: 02, January 2016, pp. 1123-1229.