International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 02 | Feb -2017 www.irjet.net p-ISSN: 2395-0072

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

Kajal K. Nandeshwar[1], Prof.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 - *Web is growing constantly and exponentially every day. Thus relevant information gathering becomes unfeasible. Web mining is nothing but the extraction of useful contents of information. Many learning object repositories stores high quality learning materials. High learning materials are expensive to create, so it is very important to ensure reuse of learning material. The learning material can be tagged either manually or automatically. Manual annotation is time consuming and expensive process. Regarding automatic gathering information system various proposals have been developed. The first is the Agathe which is multiagent system for gathering the information on restricted domain. Next one is Crossmarc which is multi domain system based on multilingual agents for extracting information from web pages. Next is CiteSeerX which is scientific literature digital library and search engine. In this paper, a review is proposed for increasing the efficiency of automatically gathering information.*

***Key Words*: EDOs, gathering information, website, repository, automatic, extraction**

1. INTRODUCTION

Among other things, as an educational information source, internet is used [1]. The learning object repository is storing content or assets or resources as well as their metadata record [2]. Educational resource also called digital object, learning object, learning resources, digital resources, digital content, reusable learning object, educational content (McGreal, 2004). An Educational Digital Object (EDO) is any material in digital format that can be used as an educational resource. For example, a scientific publication, an educational material used in a class is an educational resource [6].World Wide Web (WWW) is a vast repository of interlinked hypertext documents known as web pages. A hypertext document consists of both, the contents and the hyperlinks to related documents. Information on the Web is very huge in size. For effectively satisfying the information need of the user on the Web, there is a need to use this big volume of information efficiently [7].

There are various systems are used for automatic gathering information such are Agathe, CROSSMARK, CiteSeerX. All these works consider documents that have some type of structure such as call for papers or scientific papers. In all cases previously analyzed, only information that is contained in the document is extracted but they did not explore information that could be in linked websites [6].

2. Preliminaries

2.1 Web crawling

A Web crawler is a program which inspects web pages in a methodical and automated way [8].

One of its common 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 and analyzed in order to look for links to new resources, typically HTML pages, repeating this process until the final conditions are reached. These conditions vary according to the desired crawling policy [6].

2.2 Information Extraction, Retrieval, and Gathering

The main goal of Information Extraction systems is to locate information from text documents in natural language, producing as output a structured tabular data without ambiguity, which can be summarized and presented in a uniform way [9]. Increasingly, it is necessary to extract information for different purposes from the web [6].

The relevant documents within a larger collection of documents are retrieves by an Information Retrieval system, while relevant information in one or more documents is extracted by an Information Extraction system. Therefore, both techniques are complementary and used in combination can result in powerful tools for text processing [6].

Because of growth of the web and heterogeneity of its pages, the gathering information is increasingly complicate. For performing the retrieval and extraction of information in well-defined collections, a Gathering Information System is responsible. To retrieve relevant information, the gathering should be restricted to the specific domains [6].

3. Proposed System

The architecture of the proposed system containing a query which takes text (it can be a URL of the website) as an input and then the crawler crawling website and extracting its contents. Then EDOs are collects from the crawler. Then these are classified by the classifier which extracted contents into the audio, video, and text and these EDOs are saved to the database. Then what users needs like audio, video, and text and extract the data. The following flow chart illustrating the system architecture that collecting text 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 title, category, author, language, keywords and relevant contact data are automatically extracted.

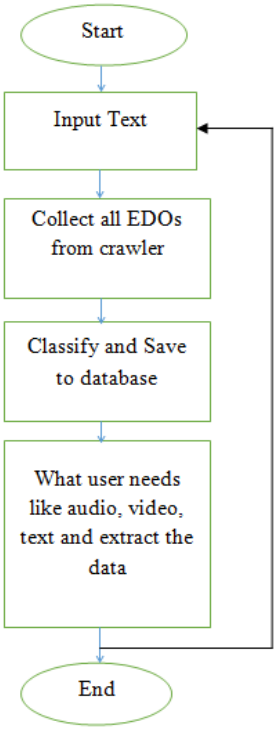

Fig -1: Flow chart of proposed system

4. Background

Various proposals have been developed for the automatic gathering information systems. Following systems are used for the automatic gathering information. The Agathe described by Albitar, Espinasse, and Fournier [7] that proposes a generic multiple agent architecture for contextual information gathering on the restricted web domains. In Agathe, the ontologies are exploiting by software agent in order to realize web page classification and information extraction tasks. Currently in Agathe, firstly the web pages are retrieved from the web. Next some agents use domain ontology to classify them semantically. Then, other agents who are depending on same ontology extract relevant information from these pages according to their classes. The AGATHE system is a generic software architecture that allowing development of information gathering systems on the Web, for one or more restricted domains [3]. The authors use ontology to consider context information to restrict the recopilation on the web to a certain domain [6].

The CROSSMARC [4] is a project of multiple domain system which support development of an agent based multilingual system for extraction of information from the web pages. It uses an approach which is based on knowledge combined with machine learning techniques for designing a robust system in order to information extracting from websites of interest. CROSSMARC reduces the high cost of maintaining the system. Because of the constant change of the web, this hybrid approach supports adaptability to a degree of independence and new emerging concepts from specific web sites considered in the training phase [6].

The CiteSeerX [5] is the next generation of 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 modular web services, pluggable service components, distributed object repositories and transaction safe processes. Its architecture enhances flexibility, scalability and performance.

Table -1: Comparative study of techniques

System	Author	Advantage	Disadvantage
Agathe	S. Albitar, B. Espinasse, S. Fournier	Enabling users to state queries, scale economy	Program size becomes bigger
Crossmarc	M. T. Pazienza, A. Stellato, M. Vindigni	Reduce high system maintenance cost	Need domain experts and knowledge engineers for task of updating ontology and lexicons of domain
CiteSeerX	H. Li. et al.	Automated information extraction, focused on crawling	Errors in extraction of authors and titles

5. CONCLUSIONS

The information gathering is more and more difficult because of growing size of the web and the heterogeneity of accessible pages. The all systems regarding to automatic gathering information such as Agathe, CROSSMARC, CiteSeerX are good reference architectures for information

gathering but in all cases, only information that is contained in the document is extracted but they did not explore information that could be in another page of the same website. In the proposed architecture the data or metadata extracted are searched in the document and are also searched into the linked sites.

REFERENCES

[1] T. Pire, B. Espinase, A. Casali, and C. Deco, "Extracción automatic de metadatos de Objetos de Aprendizaje:Un studio comparativo," in Proc.4th Congreso Tecnología Educación Educación Tecnología, Salta, Argentina, Jun. 2011, pp. 1–10.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] Espinasse, B., S. Fournier, and F. Freitas, "AGATHE: An Agent- and Ontology-Based System for Gathering Information about Restricted Web Domains", in International Journal of E-Business Research (IJEBR), 2009. 5(3): p. 14-35.

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

[9] L. Eikvil, "Information extraction from World Wide Web—A survey, "Norwegian Comput. Center, Oslo, Norway, Tech. Rep. 945,1999.