IJCTA, 9(40), 2016, pp. 425-431 ISSN: 0974-5572
© International Science Press

A Proposed Prototype for Harnessing Open Corpus Resources for eLearning

Mahantesh K. Pattanshetti*, Sanjay Jasola* and Vivek Gupta*, & Himanshu Joshi*

ABSTRACT

A majority of the technology-enabled learning (TEL) systems use content that is largely proprietary to deliver eLearning courses. But the development of these systems entails large expenses due to the huge manual effort, is time-consuming, requires expertise, burdens the faculty and despite these drawbacks may not address the unique learning requirements of all learners. These systems also suffer from two major drawbacks namely the lack of resource *reuse* and *interoperability* between eLearning systems. Adaptive hypermedia systems (AHSs) using freely available learning resources offer a potential resolution to the aforementioned issues. Although, these learning resources on the web known as *open corpus resources (OCRs)* can be tapped for eLearning they come with their own set of problems. The major impediments to learning objects reuse are locating suitable content from massive volumes of data on the web, lack of resource metadata for semantic interpretation and machine processing, difficulty in automatic sequencing and repurposing learning objects, along with maintaining pedagogical and aesthetic consistency. This paper examines the research challenges to harnessing barrier-free web content and proposes a prototype to exploit web resources for personalized eLearning.

Keywords: eLearning, open corpus resources, adaptive hypermedia systems, bridging knowledge divide, open educational resources

1. INTRODUCTION

Traditional notions of learning are being upended in the digital era. Learning today is rarely limited to the formal classrooms and textbooks. The omnipresence of the Internet coupled with the affordances of digital devices and accessibility has ensured that digital learning objects form a significant component of learning. It may not be out of order to view the web as a large digital library containing a copious array of resources, such as open educational resources (OERs), tutorials, research papers, multimedia educational objects, applications, tools, lecture notes, blogs, and online courses like massive open online courses (MOOCs) [1,2]. A significant volume of the aforementioned resources exists for personal use without any legal barriers to access. These learning resources may be accessed quickly by modern search engines like Google. The issue with search engines is that they ignore user attributes like varying background knowledge of the domain, learning styles, learning goals, and deliver "*one-size-fits-all*" results. The diverse nature of user requirements implies the need for "*just-for-me*" information delivery. Adaptive hypermedia systems (AHSs) offer a remedy to the problem of information overload, lack of personalization and user disorientation in hyperspace [3]. Common examples of AHSs in daily use are news personalization and electronic commerce.

The rest of this paper is organized to facilitate a conceptual understanding of the research domain and the issues thereof. The subsequent section commences with a primer on AHSs followed by research challenges and research motivation in section III. Review of state-of-the-art is presented in section IV. A prototype model to address the challenge of harnessing OCRs for eLearning can be found in section V, followed by concluding remarks in section VI.

* Department of Computer Science and Engineering, Graphic Era Hill University, Uttarakhand, Dehradun, India, *Emails: mahant.india@gmail.com, sjasola@yahoo.com, vivgupta95@gmail.com, himanshu234joshi@gmail.com*

2. ADAPTIVE HYPERMEDIA SYSTEMS

Hypermedia may be viewed as a mash-up of hypertext, graphics, audio, video, text, and hyperlinks. As opposed to multimedia which is static, hypermedia pages are interactive. A vast number of web pages are hypermedia pages with rich content coupled with interactivity.

By default, web search results deliver the same content irrespective of user attributes. A curative to the vanilla search results is offered by AHSs. Customization of content delivery taking into account user characteristics like knowledge, goals, preferences, learning styles, device characteristics among others is facilitated by AHSs. This adaptation effect is possible by drawing inferences from different elements as shown in figure 1. Customization of the interaction is facilitated by two mechanisms namely adaptive presentation and adaptive navigation. Examples of techniques to provide adaptive presentation may include adding or removing text, highlighting text, and dimming text. Adaptive navigation support may be provided through for example by techniques like link enabling/disabling/hiding, sorting, and link annotation to link generation [3].

Adaptation effect implementation in AHSs is facilitated by link typing to indicate the type of knowledge unit (ex: introduction, quiz, problem etc) and mapping of content with respect to a domain ontology. The user is given the impression of "*just-for-me*" content by inferring from the user model prior to content presentation. For modeling user knowledge a common approach is to model the user knowledge as a subset of the domain knowledge possessed by domain experts [3]. Prior to the commencement of a discussion on the research challenges, commonly used terminology in literature is defined.

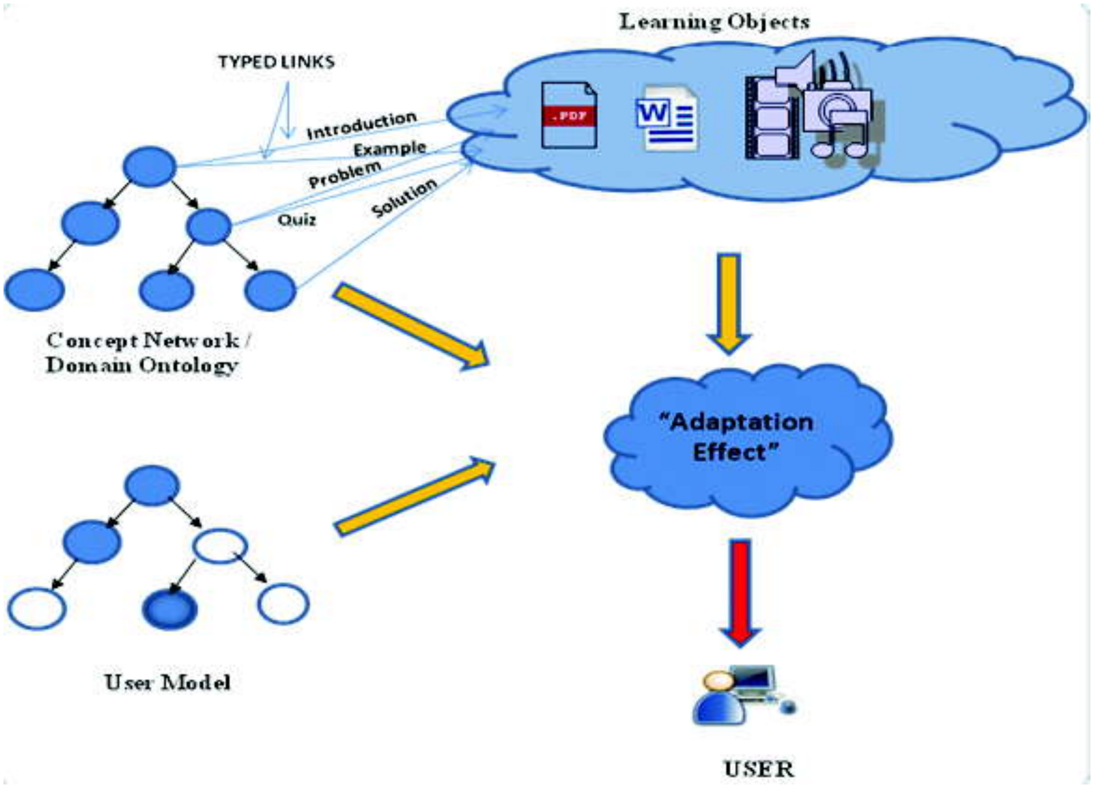


Figure 1: Conceptual view of an Adaptive Hypermedia System

2.1. Open Educational Resources (OER)

Content that is more liberally licensed, typically with creative commons license for reuse, modification, and repurposing, where just attribution is good enough without the need to seek explicit permission of the copyright holder. OER resources can be found in repositories like Multimedia Educational Resource for Learning and Teaching Online (MERLOT, www.merlot.com) [2]. In contrast, a large number of web resources come with fair use policy with very specific rights mentioned by the publisher.

2.2. Closed Corpus Resource Adaptive Hypermedia Systems

All the documents and their relationships among themselves and with respect to the reference models for adaptation are known at the time of system design. In these systems, content is proprietary and the system completely handcrafted [3].

2.3. Open Corpus Resource Adaptive Hypermedia Systems

The documents and thereby their interrelationships among themselves and with respect to reference models for adaptation is unknown at design time. In these systems, the content may be sourced from the web along with proprietary content [3].

In this paper, the focus is on constructing AHSs using open corpus resources from the web with minimal manual intervention or the requirement of sophisticated IT skills.

3. RESEARCH CHALLENGES AND MOTIVATION

Although the web may be viewed as a vast repository of learning resources, it comes with major challenges of resource sourcing and processing for utilization in eLearning applications. Some of the significant issues relevant to the development of AHSs using open corpus resources are summarized below.

- The overabundance of data on the web comes with the challenge of *locating accurate content* pertinent to a user model and user requirements [5].
- Technical challenges of natural language processing (NLP) coupled with the sparseness of*metadata* or *semantic annotation* on a significant corpus of web resources renders the application ofautomatic machine processing techniques arduous [3].
- Most of the content on the web is contributed and more relevant to the learning requirements of developed countries. Adapting content to the requirements of the developing world, especially to a country like India with numerous languages and sub-cultures, is complex, to say the least. In literature, this is referred to as the *transculturation* problem [4].
- Awareness of the potential to harness the web as a source of knowledge is close to a decade and a half, but the research topic is in very *nascent stage* as can be seen with just a handful of publications addressing the research challenge [1, 5 and 6].
- The *inter-disciplinary nature* of the domain requiring expertise in AI, IR, big data, NLP and educational theory adds an additional layer of complexity to the research problem.

India has one of the youngest populations in the world. In order to harness this demographic dividend, education is vital. Unfortunately, the country is in the midst of a national crisis in education. The poor quality of teachers is evidenced by less than one percent qualifying the entrance test conducted for teacher recruitment by the government of India. In addition, numerous reports have highlighted the falling reading and mathematical skills among students [7]. The research aims to explore the utilization of barrier free openly available digital resources for eLearning to address the falling academic standards.

Goals of the proposed research conform to the objectives of the National Knowledge Commission (NKC) which recognizes that success in a knowledge economy relies on access to high-quality education through creation and dissemination of learning resources [8]. A significant research motivation is a conviction that barriers to social mobility and knowledge divide can be addressed by providing access to *high quality* and *cost-effective* eLearning solutions.

Although digital rights, security, and intellectual property rights are significant in the web context, this paper chooses not to address them in order to avoid scope creep and maintain a singular focus on constructing eLearning systems using web resources.

4. REVIEW OF STATE OF ART

Early efforts commenced with the display of printed textbooks as digital copies. Gradually, the functionality of online courses evolved concomitantly with technical advancements in web technology. Mid 1990's can be identified as the turning point where research efforts were jumpstarted by individual researchers working in western universities. These early systems are known as *closed corpus systems* and primarily used custom developed content along with proprietary architecture to support adaptation. In a closed corpus system, the content and adaptation were strongly intertwined to facilitate learning by a target audience known in advance. As a result, their use was restricted to the audience they were developed for [9].

In the second phase, researchers made attempts to incorporate pre-selected web content along with their closed corpus offerings. These systems are referred to as *mixed corpus systems*. As an upshot, learners were able to enhance their learning outcomes by harnessing additional pre-selected web resources. This was a significant improvement over the previous systems but it still lacked flexibility to dynamically source learning resources due to the lack of semantic knowledge of web learning objects [10, 11, 12, and 13].

The third phase currently in vogue is characterized by researchers attempting to develop systems using resources exclusively from the web. Only a handful of attempts have been made so far and they come with significant restrictions and manual effort. A characteristic of these systems has been the move towards component based service oriented architecture. These *open corpus systems* have been primarily developed in the western nations and address their educational requirements [1, 5 and 6]. There is a significant scope for exploring the use of open corpus resources for eLearning in the context of the developing nations.

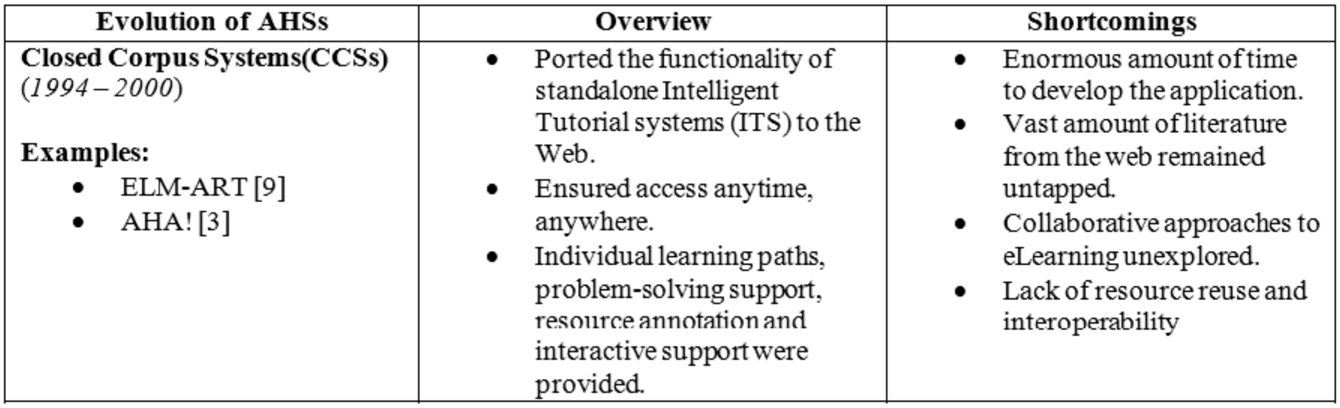
The progress of AHSs with regard to their use of learning resources along with a critical analysis is provided in table 1.

A close examination of the shortcomings of systems using only OCRs reveals that a number of research gaps exist. For one, these systems have primarily targeted tertiary education; this leaves a whole lot of informal learners and other tiers of education waiting to be served. Also, the processes and methods adopted are highly technical and laborious requiring an advanced knowledge of IT. The proposed prototype takes into consideration the drawbacks of the existing systems and offers a simpler means of facilitating eLearning using the web.

5. ARCHITECTURE OF THE PROPOSED PROTOTYPE

In order to address the challenges ranging from content sourcing to delivering content in a "just-for-me" format to the end user a pipelined processing architecture as in [14] is proposed. Each layer of processing shall contribute to the final objective of delivering learning objects to a user adhering to a suitable pedagogy

Table 1
Evolution of Adaptive Hypermedia Systems



Evolution of AHSs	Overview	Shortcomings
Closed Corpus Systems(CCSs) (*1994 – 2000*) **Examples:** • ELM-ART [9] • AHA! [3]	• Ported the functionality of standalone Intelligent Tutorial systems (ITS) to the Web. • Ensured access anytime, anywhere. • Individual learning paths, problem-solving support, resource annotation and interactive support were provided.	• Enormous amount of time to develop the application. • Vast amount of literature from the web remained untapped. • Collaborative approaches to eLearning unexplored. • Lack of resource reuse and interoperability

Mixed Corpus Systems(MCSs) (*2000 – 2005*) **Examples:** • KBS Hyperbook [10] • Knowledge Sea [11] • Knowledge Sea II [13]	• Proprietary content was linked with manually selected web resources. • Indexing using manual, social collaborative annotation or AI techniques were used to provide adaptation. • Provided all functionality of CCSs with conceptual bridges to web resources.	• Manual effort in developing the CCSs was followed by manual effort to locate and link the proprietary contents with web resources. • By providing a "*closed*" view of the web, dynamic content from the web remained unexplored.
Open Corpus Systems(OCSs) (*2006 till Date*) **Examples:** • Lawless et al., [5] • Sosnovsky et al., [1] • Kravcik et al., [6]	• Systems exclusively use learning resources from the web. • Use domain ontology for mapping resources. • Tools like federated search or crawlers or manual selection for resource sourcing. • Annotation by experts or by automated tools or by algorithms.	• Processes are complex; require expertise, are time-consuming and require significant manual intervention to develop an eLearning application. • Have not explored social mechanisms for resource annotation, discovery or for quality evaluation of content. • Have not exploited advances in pedagogy and user modeling of learners.

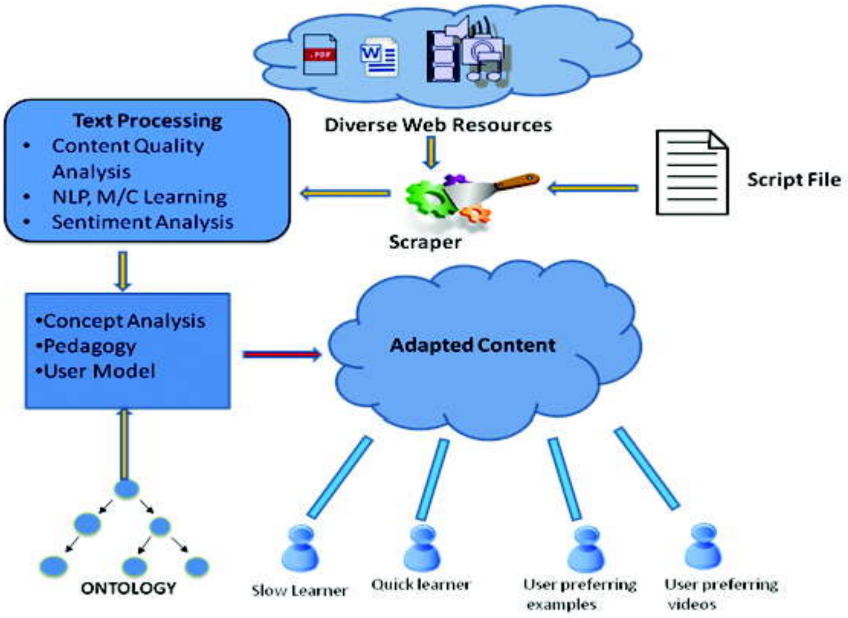


Figure 2: Anatomy of pipeline processing architecture to deliver personalized eLearning

and a user model. Figure 2 provides a high-level view of the processing steps, followed by a brief description of the pipeline modules.

- The first step is to *source* content from the web. Previous researchers have used crawlers, federated search or manual selection for resource identification. It is proposed to use a *web scraper* seeded with the scripts to locate content from the web using popular search engines. The scraper shall locate and download resources from open web plus selected repositories.
- eLearning applications for education require the highest possible quality of content. Thereby the second module shall work on screening content for high-quality learning objects. High-quality

learning objects possess certain attributes that can be discovered by application of statistical techniques. In addition certain indicators and quality dimensions are going to be examined to predict the quality of web resources.

- In this module, the extracted text shall be sanitized for textual analysis. Standard text preprocessing like language detection, word stemming, stop word removal, parts-of-speech processing, and feature construction for application of *machine learning techniques* for concept matching and relevance to the learning objectives shall be implemented.
- To ensure appropriate concepts are being presented to the learner the content shall be analyzed through concept detector applications like Alchemy API [14]. Finally, the content shall be presented to the end user taking into consideration the user model and an appropriate pedagogy.

6. CONCLUSION

The paper commenced by making a case that traditional learning models are being disrupted by the advent of digital technologies. In order to exploit advances in the digital domain and widespread availability of learning resources on the web, AHSs was shown to be suitable to facilitate personalized eLearning. Developing an eLearning system using resources available on the web comes with its own set of challenges. The primary challenge is the lack of metadata on web documents thereby making their interpretation difficult and requiring creative approaches to infer their applicability for eLearning. Reusing web learning objects for eLearning is also hampered by research being in the nascent stage as envisaged by just a handful of research papers. This lack of research progress can be attributed to a large extent to the interdisciplinary nature of the domain and also due to lack of maturity in AI techniques for natural language processing. On the flip side, a lack of substantial research and the research challenges presented offer a world of opportunity to researchers.

This paper is an immediate outcome after the research proposal has been accepted by the university research committee. The authors are currently in the process of implementing an eLearning system using the above architecture. Initial results appear promising. A comprehensive paper detailing the implementation and user evaluation shall be submitted once the trial is concluded.

ACKNOWLEDGEMENTS

The authors would like to express their profound gratitude to the chancellor of both the universities Dr. Kamal Ghanshala for providing excellent ambiance for research. In addition, the authors are also highly indebted to Dr. R.C. Joshi former Professor and Head Department of Electronics and Computer Engineering at the prestigious Indian Institute of Technology (IIT), Roorkee and currently working as an academic advisor to Graphic Era group of institutions for his invaluable suggestions and guidance.

REFERENCES

[1] Sosnovsky, S., Hsiao, I.H. and Brusilovsky, P. Adaptation "in the Wild": ontology-based personalization of open-corpus learning material. *In European Conference on Technology Enhanced Learning* (pp. 425-431). Springer Berlin Heidelberg. (2012)

[2] Butcher, N. *A Basic Guide to Open Educational Resources (OER)*. Commonwealth of Learning Vancouver and UNESCO. (2015)

[3] Brusilovsky, P., Kobsa, A. and Nejdl, W. eds. *The adaptive web: methods and strategies of web personalization* (Vol. 4321). Springer Science & Business Media. (2007)

[4] Atenas, J. and Havemann, L. Questions of quality in repositories of open educational resources: a literature review.*Research in Learning Technology*, 22. (2014)

[5] Lawless, S., Hederman, L. and Wade, V. OCCS: Enabling the Dynamic Discovery, Harvesting and Delivery of Educational

Content from Open Corpus Sources. *In 2008 Eighth IEEE International Conference on Advanced Learning Technologies* (pp. 676-678). IEEE. (2008)

[6] Kravèík, M. and Wan, J. Towards open corpus adaptive e-learning systems on the web. *In International Conference on Web-Based Learning* (pp. 111-120). Springer Berlin Heidelberg. (2013)

[7] Perryman, L.A. Addressing a national crisis in learning: open educational resources, teacher-education in India and the role of online communities of practice. *In: Seventh Pan-Commonwealth Forum on Open Learning* (PCF7), 2-6 Dec 2013, Abuja, Nigeria.,Communities of practice.

[8] Perryman, L.A., Hemmings-Buckler, A. and Seal, T. Learning from TESS-India's approach to OER localization across multiple Indian states. *Journal of Interactive Media in Education*, 2014(2).

[9] Weber, G. and Brusilovsky, P. ELM-ART–An Interactive and Intelligent Web-Based Electronic Textbook. *International Journal of Artificial Intelligence in Education*, 26(1), (2016) pp.72-81.

[10] Henze, N. and Nejdl, W. Adaptation in open corpus hypermedia. *International Journal of Artificial Intelligence in Education*, 12(4), (2001) pp.325-350.

[11] Brusilovsky, P. and Rizzo, R. Using maps and landmarks for navigation between closed and open corpus hyperspace in Web-based education"". *New review of hypermedia and multimedia*, 8(1), (2002) pp.59-82.

[12] Dolog, P., Henze, N., Nejdl, W. and Sintek, M. The personal reader: Personalizing and enriching learning resources using semantic web technologies. *In International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems* (pp. 85-94). Springer Berlin Heidelberg. (2004)

[13] Brusilovsky, P., Chavan, G. and Farzan, R. Social adaptive navigation support for open corpus electronic textbooks. *In International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems* (pp. 24-33). Springer Berlin Heidelberg. (2004)

[14] Levacher, K., Lawless, S. and Wade, V. Slicepedia: Content-agnostic slicing resource production for adaptive hypermedia. *Comput. Sci. Inf. Syst.*, 11(1), (2014) pp.393-417.