Intinno: A Web Integrated Digital Library and Learning Content Management System

Synopsis of the Thesis to be submitted in Partial Fulfillment
of the Requirements for the Award of the Degree of

Master of Technology

In

Computer Science and Engineering

Submitted By:
Arpit Jain (03CS3009)

Under the supervision of
Prof. Pabitra Mitra
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
April 2008

Abstract

The work describes the design of Intinno, an intelligent web based learning content management system. The system aims to circumvent certain drawbacks of existing learning management systems in terms of sparsity of content, lack of intelligent search and context sensitive personalization. The sparsity problem is solved by using web mining to crawl learning content from the web. The mined content is then used to automatically generate concept maps. Automatic annotation using the concept maps is used to archive the crawled content into a digital library. Multiparameter indexing and clustering is done to provide intelligent content based search. Finally, algorithms for learning applications like generation of memory maps are proposed. Context sensitive and personalized recommendation on content is supported. The system is available online at http://www.intinno.com

1. Introduction

A Learning Management System (or LMS) is a software tool designed to manage user learning processes [1]. LMSs go far beyond conventional training records management and reporting. The value-add for LMSs is the extensive range of complementary functionality they offer. Learner self-service (e.g. self-registration on instructor-led training), learning workflow (e.g. user notification, teacher approval, waitlist management), the provision of on-line learning, on-line assessment, management of continuous professional education, collaborative learning (e.g. application sharing, discussion threads), and training resource management (e.g. instructors, facilities, equipment), are some of the additional dimensions to leading learning management systems [2].

In addition to managing the administrative functions of online learning, some systems also provide tools to deliver and manage instructor-led synchronous and asynchronous online teaching based on learning object methodology. These systems are called Learning content management systems or LCMSs. An LCMS provides tools for authoring and re-using or re-purposing content as well as virtual spaces for learner interaction (such as discussion forums and live chat rooms). The focus of an LCMS is on learning content. It gives authors, instructional designers, and subject matter experts the means to create and re-use e-learning content more efficiently [3].

The current course management systems have a number of drawbacks which hinder their wide acceptance among teachers and students. One of them being the problem of cold start. Instructors who begin to make up a course don't have the material to start up. Seamless content reuse is often not possible. Materials presented may lack coverage of the subject area and thus fail to cater information needs of all students in a class. On the other hand, students while

studying or reading a lecture have to waste a lot of their time in searching for relevant resources from the web.

We aim to build a system which solves the above problems to a large extent. The web interfaced educational digital library will solve the cold start problem faced by instructors. While putting up new course, assignment or a lecture, similar resources would be available from the digital library either by search or by recommendations. Also, while reading a lecture/tutorial a student would be recommended relevant material from the web and thus would save him time spent in looking for relevant resources [4]. The system will have the additional benefit of acting as a collaboration platform among students by building up of networking of courses.

The main focus of our system (called Intinno) is to mine the free and open source material available on the web [5] to build up a quality collection of learning material. Such material is automatically annotated and archived in a digital library and later intelligently searched or recommended.

2. Architecture of Intinno

The functionalities provided by the system include, web mining, learning content management, search and recommendation. Accordingly, Intinno has the following major components:

1. Web Miner for Learning Content Discovery
2. Information Extraction and Automatic Annotation Module
3. Digital Library and Content Search Module
4. Personalization and Content Recommender System
5. Learning Applications like Automatic Book Generation, Construction of Memory Maps

A block diagram of Intinno system is shown in Figure 1. The system tasks may be mainly classified into the following major steps: Step (i) Building up of digital library from the content crawled from the web. The subtasks of this step are (a) Collection of resources from the web, (b) Knowledge Representation Architecture capturing semantics of content. Step (ii) Building applications using the Knowledge Representation Architecture. Some of the applications supported are (a) Search for similar material, namely, similar courses and similar content (lectures/ tutorials/ assignments). (b) Personalized recommendation by using the context information of current user and the current course.

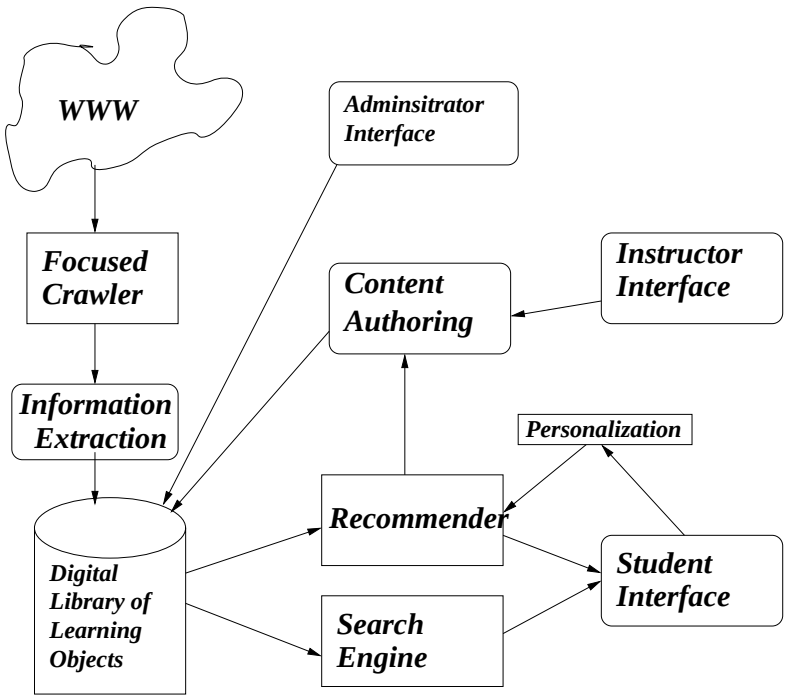

Figure 1: Intinno architecture: Block diagram

Each of the above components is described in detail in the thesis. In our work we present a survey of the popular E-learning system architectures, various data mining techniques used to implement various components. The thesis analyses the current approaches and proposes novel techniques for the use of data mining in e-learning.

3. Mining Learning Resources from Web – An Overview

Web being a rich repository of learning content, we attempt to collect high volume of learning material from web using a web miner [6]. The type of content required for the digital library would include:

1. Courses
2. Assignments
3. Lectures and Tutorials
4. Animation and Videos
5. Case Studies
6. Questions and Quizzes
7. Information of relevant technologies from the industry.

The content described above can be mined from the following major resources:

1. Websites hosting standardized, reviewed and open source course material like MIT Open Courseware, NPTEL India.
2. Course websites of large international universities. We have considered US universities currently.
3. Discussion Forums - Google Groups, Yahoo Answers
4. Websites for animations/videos - Youtube, Google Video and metacafe
5. Websites for general content - Wikipedia, Mathworld
6. Company Websites for product related info and case studies
7. Domain specific websites for questions, tutorials etc.

The resources are categorized according to the type of repository. A general strategy for crawling all the resources will not be effective. Hence we apply different crawling strategies for each category of resources. A focused crawling algorithm using sequential models [7] is implemented to mine content from the web. The approaches for crawling are discussed in detail in the thesis.

4. Domain Knowledge Representation – An Overview

The knowledge representation is crucially important in the development of an intelligent learning system for e-learning. The required documents are mined from the web as mentioned in Section 3. In order to use documents effectively and efficiently, not only the contents but also the representation of contained knowledge is important. The effectiveness of the learning applications like memory maps will depend significantly on the knowledge representation architecture. The content mined from the web can be divided into two categories in terms of its usability for an e-learning system. (i) Structured content (courses) (ii) Non-Structured educational content.

Structured course content crawled in section 3 is annotated before it is archived in the digital library. We identify a set of tags which are required to represent the course in SCORM [9] standard. We extract information from the structured courses to convert them into (semi-) SCORM format. In addition to the above tags we also store some entity specific meta-tags that are important from the point of view of indexing and parameterized search. For both the above category of tags hand crafted wrappers are used for information extraction [8]. Adding the search keywords in the meta-tags ensures that information about related course/course material is added in the tags of the entity. This will ensure that if the search is made in the name of the course then related material also turns up in the results.

Non-Structured content is available in abundance on the web. Open repositories like wikipedia and information pages authored as blogs etc:- by casual users if used efficiently can be a very good resource for learning. All this knowledge needs to be represented efficiently for use by e-learning systems. In our work, we report the existing approaches. We also propose a new approach for automatic construction of concept maps from the content mined from the web. Our knowledge representation technique is specially designed to support intelligent tutoring applications like automatic annotation of text and construction of memory maps. We have designed and implemented a heuristic based algorithm to extract the headings from web documents. We will present the architectural roadmap and results in the thesis.

5. Searching the Digital Library – An Overview

We provide two additional capabilities in addition to keyword based search, namely (i) Content based search for similar courses, and (ii) Intelligent search for course materials (i.e. if a search is given for material on biochemistry course then materials from molecular biology course should also turn up in the results.

To find similar courses we will perform hierarchical clustering on the courses since courses by nature are a hierarchical. To perform clustering on courses we need a similarity measure between different courses. A combination of measures will be used to perform clustering of courses. The clustering of courses is performed offline.

5.1 Ranking of search results

The above clustering of courses and the augmentation of tags will facilitate intelligent search of courses/material. But the search results need to be ranked too. The following criteria may be used in ranking the courses/course materials:

1. Measure indicating the authority of the source of the material (For example university from where the material is picked)
2. Cosine similarity measure on the tags by the keywords
3. Courses with more content may be ranked higher
4. For content being searched, content from similar courses may be given more preference.
5. Also content from courses or even courses with same difficulty level as the user may be given more preference.

6. Personalized Recommendation – An Overview

Students studying a particular page (content) will be recommended similar contents. This recommendation will be:

1. Content Based i.e. Content from the digital library similar to the one being read will be recommended.
2. Content Diversified: i.e., if the person is currently studying Lectures then he/she will be recommended questions/quizzes/Assignments. However if the person is bust doing Assignments or solving questions then he/she will be recommended lectures/tutorials on that particular subject from the digital library.

Recommendation will based on a learning approach where for some content the actual goodness of the link would be learned by the number of clicks on the recommended links. Recommendation will be personalized i.e it will be based on the courses that the user has done and also on his/her level of understanding which can be judged from his courses list.

7. Learning Applications – An Overview

Most of the research in the application of Data Mining for E-Learning systems has been on evaluation of the learning process and then re-organizing it according to the users needs. We present a survey of the applications in development or already being used by popular E-learning systems like Blackboard, Moodle. In our work, we propose the architecture for applications like automatic text annotation and memory maps that will be developed over the knowledge representation architecture mentioned in Section 4.

8. Implementation and Deployments

The Intinno system [9] may be used online at http://www.intinno.com. Currently there are about ten courses archived in the digital library, more are being added soon. The system currently crawls about 1000 university sites in .edu domain in addition to well known educational material resources. It also mines a number of hand picked websites.

Intinno is under use for the undergraduate courses for all departments at Indian Institute of Technology Kharagpur. User feedbacks are being collected for further improvement of the system.

9. Conclusion and Future Work

The philosophy behind Intinno was to use web mining techniques to develop an intelligent education portal. Key features include, web crawling, avoidance of cold start, automatic annotation, sophisticated indexing, intelligent search and personalized recommendation of content.

The system is under development and machine learning and data mining techniques are explored to improve the quality of user experience. Quantitative evaluation of search and recommendation performance is also being carried out.

10. References

1. *Wikipedia*: http://www.wikipedia.com

2. J. Cole and H. Foster, *Using Moodle: Teaching with the Popular Open Source Course Management System* (O'Reilly Media Inc., 2007).

3. V. B. Devedzic, *Key Issues in Next Generation Web Based Education*, IEEE Trans. System Man and Cybernetic: Part C, Vol. 33, pp. 339 (2003).

4. P. Dolog, N. Henze, W. Nejdl and M. Sintek, *Personalization in distributed e-learning environment*, in WWW04: Proc. Intl. Conf. World Wide Web, 2004.

5. D. Bergmark, *Collection synthesis*, in JCDL '02: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries, (ACM, New York, NY, USA, 2002).

6. S. Chakrabarti, *Mining the Web: Discovering Knowledge from Hypertext Data* (Morgan-Kauffman, 2002).

7. V.G.Vinod Vydiswaran and Sunita Sarawagi, *Learning to extract information from large websites using sequential models* (In COMAD, 2005. SIGKDD Explorations. Vol. 6, Issue 2 - Page 66)

8. N. Kushmerick, *Gleaning the Web*, IEEE Intelligent Systems, Vol. 14, 20 (1999).

9. SCORM: http://en.wikipedia.org/wiki/SCORM

10. http://www.intinno.com