A Generalized Knowledge Representation Architecture for Intelligent Tutoring 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:
Mayank Jain (03CS3010)

Under the supervision of

Prof. Pabitra Mitra

Department of Computer Science and Engineering

Indian Institute of Technology Kharagpur

April 2008

Abstract

*An **intelligent tutoring system** (ITS) is a system that provides direct customized instruction or feedback to students without the intervention of human beings. Most of the research in ITS has been in evaluation, restructuring of the learning objects, and personalization of the learning object navigation path. With the explosion of content on the World Wide Web(WWW), the scope of application of Data Mining to E-Learning applications has increased tremendously. In this work, we identify a set of applications which go one step ahead from ITS and use the WWW to aid the learning process of the "learning object content". Each application has a high level of coupling with the knowledge representation model, which models the resources stored in the Digital Library. This domain model must be complete and accurate otherwise the learning system will not perform effectively.*

*This research presents the architecture for simplifying and automating the process of creating the domain model for an intelligent e-learning system. In this work we present an algorithm to create a **knowledge representation** of educational resources using the World Wide Web. We outline the advantages and limitations of this general architecture, and describe its implementation in Intinno – a developing Intelligent Learning Management System.*

1. Introduction

One of the hottest R&D topics in recent years in the AI community, as well as in the Internet community, is *the Semantic Web*. It is about making the Web more understandable by machines [1]. It is also about building an appropriate infrastructure for intelligent agents to run around the Web performing complex actions for their users [2]. In order to do that, agents must retrieve and manipulate pertinent information, which requires seamless agent integration with the Web and taking full advantage of the existing infrastructure (such as message sending, security, authentication, directory services, and application service frameworks) [3]. Furthermore, Semantic Web is about explicitly declaring the knowledge embedded in many Web-based applications, integrating information in an intelligent way, providing semantic-based access to

the Internet, and extracting information from texts [4]. Ultimately, Semantic Web is about how to implement reliable, large-scale interoperation of Web services, to make such services computer interpretable to create a Web of machine-understandable and interoperable services that intelligent agents can discover, execute, and compose automatically [5].

The problem is that the Web is huge, but not smart enough to easily integrate all of those numerous pieces of information from the Web that a user really needs. Such integration at a high, User-oriented level is desirable in nearly all uses of the Web. Today, most Web information is represented in natural-language; however, our computers cannot understand and interpret its meaning. Humans themselves can process only a tiny fraction of information available on the Web, and would benefit enormously if they could turn to machines for help in processing and analyzing the Web contents [6]. Unfortunately, the Web was built for human consumption, not for machine consumption - although everything on the Web is *machine-readable*, it is not *machine-understandable* [7]. We need the Semantic Web to express information in a precise, machine-interpretable form, ready for software agents to process, share, and reuse it, as well as to understand what the terms describing the data mean. That would enable Web-based applications to interoperate both on the syntactic and semantic level. The explicit representation of the semantics of data, accompanied with domain theories (that is, ontology), will enable a Web that provides a qualitatively new level of service - for example, intelligent search engines, information brokers, and information filters [8].

There is research on important issues related to the development of the Semantic Web, and their implications for Web-based teaching and learning [9]. It describes what it means precisely to create, to find, and to use educational resources on the Semantic Web pages, as opposed to doing it on today's Web. Our work presents the background and context for activities of developing Semantic Web-based educational systems, indicates some existing applications and tools, and introduces some applications which can enhance learning using the semantics.

3. Problem Definition

The knowledge is crucially important in the development of an intelligent tutoring system for e-learning. For this work, we assume that we have a repository of educational documents mined from the web. The content described above can be mined from the following major resources (i) MIT Open Courseware, NPTEL India (ii) .edu domain (iii) Discussion Forums - Google Groups, Yahoo Answers (iv) YouTube, Google Video and Metacafe (v) Wikipedia, MathWorld (vi) Company Websites for product related info and case studies (vii) Domain specific websites for questions, tutorials etc. 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.

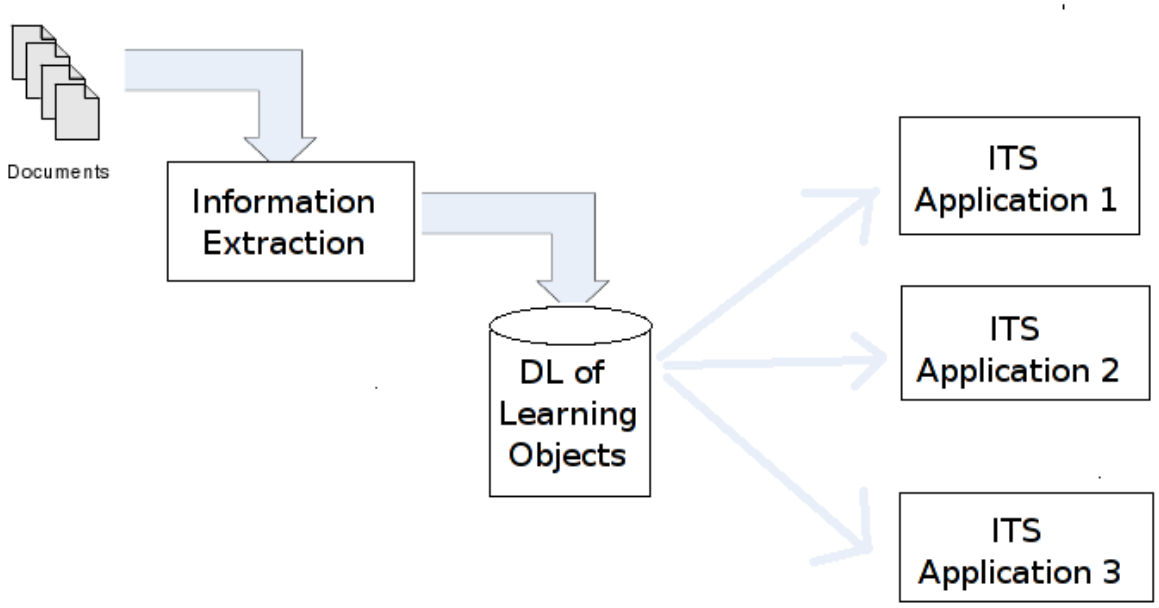
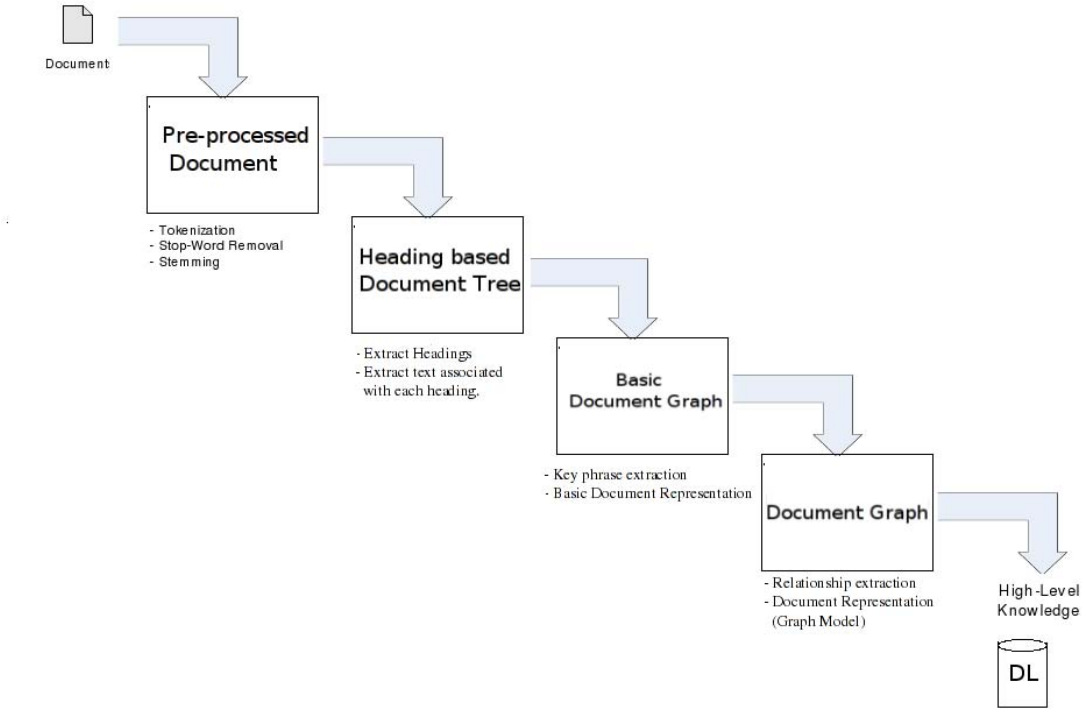


Figure 1: Problem Definition

The goal of this work is to explore approaches for representation of knowledge for efficient use of resources for an intelligent learning system. We intend to find an approach which can help in capturing the semantics of the crawled resources and efficiently implement a set of learning applications. Hence the final goal is to have a knowledge representation technique specially designed to support intelligent tutoring applications like automatic annotation of text and construction of memory maps.

4. Knowledge Representation process

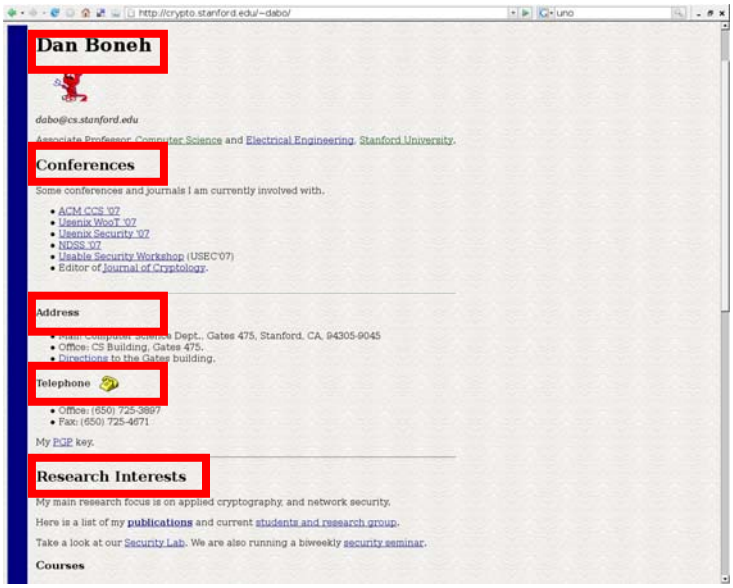
There are representation techniques such as frames, rules and semantic networks which have originated from theories of human information processing. Since knowledge is used to achieve intelligent behavior, the fundamental goal of knowledge representation is to represent knowledge in a manner as to facilitate inferences (i.e. drawing conclusions) from knowledge. Problem Solving can be simplified by an appropriate choice of *knowledge representation*. Representing knowledge in some ways makes certain problems easier to solve.



KR is most commonly used to refer to representations intended for processing by computers, and in particular, for representations consisting of explicit objects (the class of all humans, or Ram a certain individual), and of assertions or claims about them ('Ram is a human', or 'all humans have one head'). Representing knowledge in such explicit form enables computers to draw conclusions from knowledge already stored ('Ram has one head'). In our thesis, we present the detailed process of knowledge extraction using the process above.

4. Heading Extraction Algorithm Design and Implementation - Overview

We breakup the problem of heading extraction into two parts (i) Finding the headings (ii) Getting the data under the headings.



We have developed a heuristic based algorithm to extract the headings from html documents. The above algorithm was successful in extracting information from only 65% of the pages classified as faculty pages. The main problems encountered are (i) Ill-formatted HTML (ii) No Heading in the page (iii) Homepages constituting of multiple pages.

5. Ontology Generation Algorithm

We present in our thesis, a (semi) automatic framework that aims to produce a domain concept map (DCM) from text and to derive an ontology of the world from this concept map. This methodology is particularly aimed at the educational field because of the need of such structures (Ontology and CM) within the e-Learning communities to sustain the production of e-Learning resources tailored to learner's needs. Our thesis presents the detailed steps that transform textual resources (and particularly textual learning objects) into a domain concept map and explains how this abstract structure is transformed into more formal domain ontology.

This algorithm aims at bridging the gap between e-learning standard learning objects and Intelligent Tutoring Systems.

6. Keyphrase and Relationship Extraction Algorithm

There has been lot of research on keyphrase extraction from documents. The model presented by Hammouda and Kamel [10] for document representation is called the Document Index Graph (DIG). This model indexes the documents while maintaining the sentence structure in the original documents. This allows us to make use of more informative phrase matching rather than individual words matching. A list of matching phrases between two documents is computed by intersecting the sub graphs of both documents. The approach by Gutwin and Nevill-Manning [11] shows a simple procedure for key phrase extraction based on the naive Bayes learning scheme performs comparably to the state of the art. It goes on to explain how this procedure's performance can be boosted by automatically tailoring the extraction process to the particular document collection at hand. In our thesis, we present a detailed analysis of the keyword extraction process and present our conclusions.

7. Conclusion and Future Work

An efficient Knowledge representation is very essential for designing and implementing e-learning applications. More background research needs to be done to explore keyword extraction techniques. The algorithm proposed above for ontology creation needs to be refined further.

8. References

[1] Heflin, J., & Hendler, J. (2001). A Portrait of The Semantic Web in Action. *IEEE Intelligent Systems* 16(2), 54-59.

[2] Hendler, J. (2001). Agents and the Semantic Web. *IEEE Intelligent Systems* 16(2), 30-37.

[3] Scott Cost, R., Finin, T., Joshi, A., Peng, Y., Nicholas, C., Soboroff, I., et al. (2002). ITtalks: A Case Study in the Semantic Web and DAML+OIL. *IEEE Intelligent Systems* 17(1), 40-47.

[4] Gómez-Pérez, A., & Corcho, O. (2002). Ontology Languages for the Semantic Web. *IEEE Intelligent Systems* 17(1), 54-60.

[5] McIlraith, S.A., Son, T.C., & Zeng, H. (2001). Semantic Web Services. *IEEE Intelligent Systems* 16(2), 46-53.

[6] Noy, N. F., Sintek, M., Decker, S., Crubézy, M., Fergerson, R.W., & Musen, M.A. (2001). Creating Semantic Web Contents with Protégé-2000. *IEEE Intelligent Systems* 16(2), 60-71.

[7] Lassila, O. (1998). Web Metadata: A Matter of Semantics. *IEEE Internet Computing* 2(4), 30-37.

[8] Decker, S., Melnik, S., van Harmelen, F., Fensel, D., Klein, M., Broekstra, J., Erdmann, M., & Horrocks, I. (2000, September/October). The Semantic Web: The Roles of XML and RDF. *IEEE Internet Computing* 4, 63-74.

[9] Vladan Devedzic, Education and the Semantic Web. *International Journal of Artificial Intelligence in Education* 14 (2004) 39-65

[10] Hammouda, K., & Kamel, M. (2006). Data mining in e-learning. In Samuel Pierre (Ed.), E-learning networked environments and architectures: A knowledge processing perspective, *Springer Book Series: Advanced information and knowledge processing (pp. 1 – 28).*

[11] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning. 1999. Domain-specific keyphrase extraction. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence.*