Towards a Standardized Representation of Syllabi to Facilitate Sharing and Personalization of Digital Library Content

Manas Tungare[1], Xiaoyan Yu[1], Manuel Pérez-Quiñones[1], Edward A. Fox[1], Weiguo Fan[1], and Lillian Cassel[2]

[1] Virginia Tech
Blacksburg, VA, USA
{manas, xiaoyany, perez, fox, wfan}@vt.edu
[2] Villanova University
Villanova, PA, USA
lillian.cassel@villanova.edu

Abstract. A course offering is a collection of learning objects composed together based on a syllabus. Syllabi define the contents of the course, as well as other information such as resources and assignments. Currently, there is no standard format for representing syllabi that can facilitate automatic processing of syllabi contents for various applications. In this paper, we report on the current practices in creating and publishing syllabi and present the motivation for a standardized syllabus schema. We report on our experiences obtaining and identifying syllabi published online by various institutions, and extracting syllabus data from them using genetic algorithms and other machine learning techniques. Finally, we describe the tools needed for working with syllabus schema, and applications that will be made possible with the availability of syllabi in standardized formats.

1 Introduction

A syllabus forms the backbone of a course offering: a complete syllabus typically includes the course number, title, a description, the learning objectives of the course, a list of the topics covered, links to reference material such as books or publications, and other related information. The various learning objects that are included in a course offering are created based on the syllabus definition, and are tightly integrated with the reference material (also included in the syllabus).

Thus, knowledge of a course syllabus can be used to assess the structure of a course, the exact knowledge units covered, and the relative time devoted to each of them; i.e., the syllabus describes how individual learning objects are combined to form larger entities and packaged as a course for students. Although the contents of a learning object for a particular subtopic may be more or less similar at various institutions, the syllabi for similar courses at distinct institutions exhibit differences.

We are researching syllabi to better understand their role in providing information about courses. In this paper, we discuss the motivation behind investigating syllabi, various ways in which syllabus information can be made accessible in a standardized format

for specialized queries, approaches to collect syllabi published at various institutions, and tools that can help with managing and understanding syllabi. We also discuss the possible applications of a syllabus repository and the potential use of such a collection in personalizing digital library content to offer relevant additional reference material to instructors and students as they use course management systems in their everyday lives.

1.1 How Syllabi are Published Today

Many institutions have online copies of syllabi for their courses. However, there is high variation in the details of how syllabi are published. The differences can be summarized as below:

– **Publication:** Syllabi may be published by an instructor on his/her website, or via a course management system. In certain instances, syllabi are not published online, but only made available to students on printed paper.
  Most universities maintain a standardized course catalog that contains syllabus definitions for all courses offered at that school. Some of them include course summary information in the catalog, while the details are only available in university records systems. This information, however, is not in a standardized format across all universities, and usually contains only the most basic course information without details of learning objects.
– **Format:** Popular formats for online publishing include HTML and PDF. Some instructors may also post syllabi as word processor documents such as Word.
– **Access:** Some course management systems and network administration policies may prevent syllabi from being viewed outside the internal computer network of that particular institution. However, efforts are currently underway to make syllabi and course content available online licensed under an open license. The Massachusetts Institute of Technology has taken steps towards such a system (MIT OpenCourseWare) [1].

1.2 Constituents of a Syllabus

Most syllabi contain common sections. Some of these may be absent in certain syllabi, and some other unique sections may be present too, but a majority of syllabi include the following:

– Course Title
– Course Description
– Information about:
  • Instructors
  • Teaching Assistants
– Course Prerequisites
– Topics Covered
– Knowledge Units Covered (within the given Topics)
– Learning Objectives
– Course Calendar

- Reading List
  - Books
  - Book Chapters
  - Articles and Papers
  - Other Scholarly Publications
- Classroom Material
  - Presentation Slides
  - Instructor's Notes
  - Assignments

We are attempting to automatically parse this information from available syllabus documents in order to create standard representations of syllabi.

1.3 Reuse of Syllabi

Most course management systems (CMSs) include minimal built-in support for reuse of syllabi. Open-source CMSs such as Moodle and even commercial ones such as Blackboard support exporting and importing entire courses from a previous offering. The main limitation with this approach is that the export or import feature works only with the (possibly proprietary) format associated with that particular CMS. There is no industry standard for exchange of this information.

There also is no standard to help tease out the syllabus information from the course material. Connections between the syllabus and parts of the course are not explicitly indicated in current systems. Syllabi cannot be split into subparts based on either a description or a list of the learning objects. Thus, syllabi cannot be reused independent of the course material.

2 Previous Work

Some have addressed the problem of lack of standardization of syllabi. Along with a defined syllabus schema, SylViA [2] supports a nice interface to help faculty construct their syllabi in a common format.

Matsunaga *et al.* [3] describe a syllabus acquisition approach similar to ours, but it differs in the way the syllabi are identified. They crawled web pages from Japanese universities and sifted through them using a thesaurus with common words occurring often in a syllabus. A decision tree was used to classify syllabus pages and entry pages as a syllabus (for example, a page containing links to all the lectures in a particular course). Syllabus Finder [4] also helps users look for syllabi on a specific topic by performing a Google search behind-the-scenes.

Our work also relates to genre detection and metadata extraction. Research in genre detection aims to classify data according to genre, by selecting features that distinguish one genre from another, for example, identifying home pages [5]. Much work has been done on metadata extraction from papers. For example, Han *et al.* [6] describe a Support Vector Machine (SVM) classification-based method for metadata extraction from a paper's header field. Takasu *et al.* [7] describe bibliography field extraction using a Hidden Markov Model [8] (HMM).

3 Obtaining Syllabus Information

We believe that it is not likely that a syllabus schema would be adopted and used widely by academia in a short period of time. However, in order to take advantage of a schema, we must have access to a sizeable number of syllabi in a short while. Given the extensive availability of syllabi as published web documents, we decided to gather them automatically, and parse them into fields aligned with our schema definition. Any errors in such automatic parsing can be corrected via Syllabus Editors (discussed in Section 4.1) or with the assistance of the community (discussion in Section 4.3).

3.1 Obtaining a Document Set

We used the popular Web search engine, Google, to locate documents that are highly likely to be syllabi. This was accomplished by a set of two specialized queries to Google: the first, to locate departments of interest within an educational institution, and second, to locate syllabi within that department. Since our current interest is in accumulating syllabi in the field of Computer Science, we issued a query for

```
“computer science site:edu”
```

The result of this query was a list of pages within the computer science departments of many university websites (i.e., those whose domain names end with ‘.edu’). We plan to later extend our queries to include departments named, for example, ‘Computing Sciences’ or ‘School of Information Sciences’. We processed each URL to obtain the relevant domain name, i.e., the homepage of each institution. We then issued the second set of queries, one per institution, to locate syllabi within that department. As an illustration, the following query resulted in a list of syllabi from the Department of Computer Science at Virginia Tech.

```
“syllabus site:cs.vt.edu”
```

The syllabus documents thus obtained were stored in a structured organization (for later retrieval), and converted from their native format (either PDF, PS or HTML) to plain text to enable further text processing.

3.2 Automatic Syllabus Identification

The simple inclusion of a page within Google's search results is not a definitive indication of a particular webpage being a syllabus. From our study of the 7900 documents obtained, and preliminary work on a syllabus schema, we categorize them into 4 types: full syllabus, partial syllabus, syllabus entry, and noise. A full syllabus is one that contains most of the basic syllabus components and no links to other syllabus components. A partial syllabus contains some important syllabus components along with links to other components. A syllabus entry page, (for example, a course web site home page), contains a link to a syllabus, or to the various pieces that make up a complete syllabus. The rest are noise where the keyword ‘syllabus’ might occur several times, such as some articles about how to write a syllabus.

We will select content and form features to distinguish among these 4 categories. The content features include frequency of key words and phrases in a file. A thesaurus will be built containing important words and variations of them such as 'instructor' and 'course leader', 'teaching assistant' and 'grader', and 'course description' and 'course summary'. The form features include the type of document, the URL of the document, the positions of some important words such as 'syllabus' in the document, the number of links and their positions relative to important words in the document (if it is an HTML document), and so on. We have already labeled 1000 documents as one of the 4 types. With the extracted features and the 1000 documents as a training set, we can train a syllabus classifier using a machine learning approach such as the decision tree algorithm [9], support vector machines [10] (SVM), or naïve Bayes methods [9].

3.3 Metadata Extraction

Given a syllabus, our next step is to extract important components from it such as the course title, course description, instructor, schedule and topics. A syllabus can be viewed as a collection of metadata for a course. Therefore, machine learning approaches like support vector machines [10] (SVM) and hidden Markov models [8] (HMM) also can be applied since they work well to extract metadata. They also have been used to extract limited metadata from educational resources such as syllabi [11].

4 Tools for Working with Syllabus Schema

Currently, the tools for creating a syllabus document are mostly word processors or limited web forms in course management systems. Since there is no defined semantic structure, the layout and presentation capabilities of a word processor are enough to create a syllabus document. However, with the establishment of a syllabus schema, such general purpose tools may no longer be adequate to satisfy the needs of an instructor, and specialized tools will be needed. In addition to simply aiding the creation of syllabi, such tools will promote publishing syllabi to a wider audience. Some universities already use specialized tools internally [2] for creation of their own syllabi.

4.1 Syllabus Editors

We are in the process of creating a specialized Syllabus Editor to help create and edit syllabi. The use of tools, we believe, will encourage quicker adoption of the schema. We also realize the value of integrating a syllabus editor into course management systems so that the use of specialized tools will be seamless from the instructors' point of view.

Our Syllabus Editor will make it easy to announce the presence of newly-created syllabi to interested parties (via RSS/Atom feeds) and enable the inclusion of the syllabus definition in a Syllabus Repository (depending on the licensing options chosen by the instructor/institution).

4.2   Syllabus Repository

Although syllabi created at individual institutions will be hosted locally for the benefit of students, there also is value in collecting syllabi at a central repository. Features that such a repository would make possible include a search engine specifically for syllabus information, visualizations that make use of the enormous number of syllabi available, as well as other statistical analyses that heavily benefit from the presence of large numbers of syllabi.

We are exploring the possibility of loading our syllabus collection into SIMILE (**S**emantic **I**nteroperability of **M**etadata and **I**nformation in un**L**ike **E**nvironments) [12] which extends DSpace [13] so that users can navigate to and contribute to syllabi along with other educational resources in a seamless way. SIMILE will leverage and extend DSpace, enhancing its support for arbitrary schemata and metadata, primarily though the application of RDF and semantic web techniques [12].

4.3   Community-Assisted Classification and Error Correction

It is likely that syllabi classified and parsed automatically will have some errors. Most of these can be spotted by users of the system, and leveraging the assistance of the community would be helpful to maintain the quality of the repository. Success (to a high degree) has been achieved in open community-based systems in creating knowledge-bases [14] and in correcting structured information that was automatically parsed by autonomous agents [15] (referred to as Distributed Error Correction.)

4.4   Linking to the Computing Curricula Project, 2001

The Computing Curricula Project [16] (CC 2001) was undertaken to develop guidelines for computing curricula at an undergraduate level. In its final report, the authors present detailed coverage of the CS body of knowledge, core areas for undergraduate studies, learning objectives and curriculum models. In particular, the report recommends the number of hours to be devoted to particular knowledge units for courses with specified learning objectives as a guideline to instructors preparing a course.

This data can be of use to instructors when creating a syllabus, either to design their course based entirely on CC 2001 recommendations, or to pick the right mix of knowledge units, given their unique goals for a particular course.

Our Syllabus Editor will have the option of selecting knowledge units from CC 2001 to 'tag' a particular syllabus. This would allow instructors creating courses to use the content and knowledge included in the CC 2001 at course creation time. The Syllabus Editor will also include a browser of the knowledge units and brief descriptions included in the CC 2001 for each unit.

5   Applications

Availability of a standardized schema for syllabi opens up possibilities for many innovative applications. Although we envision and describe a few of them in this paper, the presence of semantically-tagged syllabus information will almost certainly lead to new ways of using that information.

5.1 Personalizing NSDL Content for Students

The National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL), funded by the National Science Foundation, has the potential to have a significant impact on the future of education in this country and around the world. We are exploring how to best tailor NSDL functionality toward particular communities of users, including learners and educators. We believe that, for the NSDL to have its intended impact, it needs to be integrated into the current pedagogical practices of educators and students.

With syllabus information available for courses, we will be able to recommend resources to students automatically, based on their skill level, and the focus of the course. Classification of the syllabus according to standardized syllabus classification schemes (discussed earlier in this paper) will help identify the relative importance assigned by the instructor to the knowledge units covered, thus enabling us to provide a proportionate amount of NSDL resources for further exploration. Previous efforts in presenting personalized content to users include the recommendation system within CiteSeer [17].

To personalize the content offered via Course Management Systems, we plan to use techniques from genetic programming. ARRANGER (**A**utomatic **R**endering of **RAN**king functions by **GE**netic programming) [18] is a discovery engine developed by one of the co-authors. We will use ARRANGER to personalize NSDL content available to users. Given a set of documents along with its relevance information from a user, ARRANGER can automatically tune a ranking function based on syntactical and lexical evidence embedded in the documents and can discover a personalized ranking function that can be used to reorder information based on personal preferences. Combined with effective user profiling and/or user feedback, ARRANGER can deliver even higher quality information to end users.

5.2 Assisting Instructors when Creating New Syllabi

When an instructor creates a new syllabus, there is a high likelihood of the end-product being similar to other syllabi (or perhaps a combination of two or more syllabi.) Although the similarity between two syllabi at the same institution is likely to be minimal, the availability of syllabi from other institutions increases the chances of finding a match. Recommendations can be made to the instructor based on resources from similar syllabi to assist with one or more of the following tasks:

– Recommending the relative number of hours to be assigned to each knowledge unit, as per standardized syllabi classification schemes.
– Locating books, reading material, and other resources that similar syllabi have included.
– Enabling the instructor to import material from other syllabi (licensed appropriately) instead of having to recreate the entire syllabus involving extra effort.

5.3 Syllabi Overview for Students

With the availability in a standardized format of all the courses that a student plans to take in a given semester, he/she can see an overview, or run specialized queries against

all syllabi at once. For example, ready answers can be obtained to questions such as "show me the complete list of books that I will need to buy this semester."

5.4 Assisting Curriculum Design, Assessment and Accreditation

With a complete definition of syllabi available, it will be easy to visualize the distribution of topics (or Knowledge Units from CC 2001) across a series of related courses. The topics used in a course can easily be used to compare two different courses and get a sense of if they are similar or not based solely on the matching of knowledge units covered in them. This could be done automatically or manually.

Also, for accreditation purposes, the information contained in the syllabus regarding learning objectives could be used to map it to program outcomes. This would assist instructors in creating a mapping of courses to outcomes, thus helping the creation of assessment processes and institutional research programs.

5.5 Comparing Programs Offered by Different Schools

There are bound to be subtle differences between syllabi at various schools. In case of multi-disciplinary topics such as Human-Computer Interaction (HCI), a particular school may lean towards a particular area while devoting lesser class time and resources to other areas. We could use a collection of syllabi from a particular university and get a sense of the emphasis given at that institution to the different sub-areas within computing. This will be helpful to students to obtain better information about specific programs in computer science before enrolling in one that fits his/her interests more closely.

5.6 Discovery of Resources

Each syllabus record will contain details of recommended resources to assist with study. Enthusiastic students will be able to turn to similar syllabi from other schools to discover additional resources easily.

5.7 Supporting Mobile Access to Educational Resources

Semantic descriptions of syllabi make it possible to personalize that information for access from mobile devices such as PDAs and cellphones. Given the intrinsic limitations of portable devices in terms of screen size, processing power, battery life and input/output capabilities, the information in a syllabus definition must be modified before presentation to a mobile client.

The availability of contextual cues during a user's interaction with a mobile device (for example, time or location) can be used to tailor the view of the syllabus, making it different from the view of the same syllabus on a standard desktop-sized computer.

6 Conclusion

In our work, we want to leverage the role a syllabus plays in defining various aspects of a course. The lack of standardization of a semantic description of syllabi has resulted in instructors and institutions publishing their syllabi in a wide variety of formats. We are working towards creating such a standardized schema, and are developing techniques to obtain and analyze the vast numbers of syllabi already available online in unstructured formats. The creation of such a repository will enable novel applications, some of which we discussed in detail. We hope that the adoption of such a schema by the academic community will help create new possibilities for instructors, students, and other entities involved in the learning process.

7 Acknowledgments

References

1. Massachusetts Institute of Technology: MIT OpenCourseWare. http://ocw.mit.edu/ (2006)
2. de Larios-Heiman, L., Cracraft, C.: SylViA: The Syllabus Viewer Application. http://groups.sims.berkeley.edu/sylvia/ (2006)
3. Matsunaga, Y., S., Y., Ito, E., Hirokawa, S.: A web syllabus crawler and its efficiency evaluation. In: Proc. ISEE. (2003)
4. Center for History and New Media, George Mason University: Syllabus Finder. http://chnm.gmu.edu/tools/syllabi/ (2006)
5. Kennedy, A., Shepherd, M.: Automatic identification of home pages on the web. In: Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05). Volume 04., IEEE Computer Society, IEEE Computer Society (2005)
6. Han, H., Giles, C.L., Manavoglu, E., Zha, H., Zhang, Z., Fox, E.A.: Automatic document metadata extraction using support vector machines. In: JCDL '03: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital Libraries, Washington, DC, USA, IEEE Computer Society (2003) 37–48
7. Takasu, A.: Bibliographic attribute extraction from erroneous references based on a statistical model. In: JCDL '03: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital Libraries, Washington, DC, USA, IEEE Computer Society (2003) 49–60
8. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. (1990) 267–296
9. Mitchell, T.: Machine Learning. McGraw-Hill (1997)
10. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc., New York, NY, USA (1995)
11. Thompson, C.A., Smarr, J., Nguyen, H., Manning, C.: Finding educational resources on the web: Exploiting automatic extraction of metadata. In: In Proc. ECML Workshop on Adaptive Text Extraction and Mining. (2003)
12. The SIMILE Project: The SIMILE Project. http://simile.mit.edu/ (2006)
13. DSpace: DSpace.org. http://www.dspace.org/ (2006)
14. The Wikimedia Foundation: Wikipedia.org. http://www.wikipedia.org/ (2006)

15. Lawrence, S., Bollacker, K., Giles, C.L.: Distributed error correction. In: DL '99: Proceedings of the fourth ACM conference on Digital libraries, New York, NY, USA, ACM Press (1999) 232

16. The Joint Task Force on Computing Curricula: Computing Curricula 2001. Journal on Educational Resources in Computing (JERIC) **1**(3es) (2001) 1

17. Bollacker, K.D., Lawrence, S., Giles, C.L.: A System for Automatic Personalized Tracking of Scientific Literature on the Web. In: DL '99: Proceedings of the Fourth ACM Conference on Digital Libraries, New York, NY, USA, ACM Press (1999) 105–113

18. Fan, W., Fox, E.A., Pathak, P., Wu, H.: The effects of fitness functions on genetic programming-based ranking discovery for web search: Research articles. Journal of the American Society for Information Science and Technology **55**(7) (2004) 628–636