Chapter IV

Automatic Syllabus Classification Using Support Vector Machines

Xiaoyan Yu
Virginia Tech, USA

Yubo Yuan
Virginia Tech, USA

Manas Tungare
Virginia Tech, USA

Manuel Pérez-Quiñones
Virginia Tech, USA

Weigo Yuan
Virginia Tech, USA

Edward A. Fox
Virginia Tech, USA

ABSTRACT

Syllabi are important educational resources. Gathering syllabi that are freely available and creating useful services on top of the collection presents great value for the educational community. However, searching for a syllabus on the Web using a generic search engine is an error-prone process and often yields too many irrelevant links. In this chapter, we describe our empirical study on automatic syllabus classification using support vector machines (SVM) to filter noise out from search results. We describe various steps in the classification process from training data preparation, feature selection, and classifier building using SVMs. Empirical results are provided and discussed. We hope our reported work will also benefit people who are interested in building other genre-specific repositories.

INTRODUCTION

A course syllabus is the skeleton of a course. One of the first steps taken by an educator in planning a course is to construct a syllabus. Later, a syllabus can be improved by adding updated course information.
or borrowing information from other relevant syllabi. Students prepare for a course by reading a course syllabus to identify textbooks. Students may use the syllabus to identify course policies, assignment deadlines, etc., during a school semester. Typically, a syllabus sets forth the objectives of the course. In addition, a life-long learner identifies basic topics of a course and popular textbooks by comparing syllabi from different universities. A syllabus is thus an essential component of the educational system.

Free and fast access to a collection of syllabi could have a significant impact on education. Unfortunately, searching for a syllabus on the Web using a generic search engine is an error-prone process and often yields too many irrelevant links. As an alternative, the MIT OpenCourseWare project (Hardy, 2002), which provides free access to MIT course materials, is a good start towards making a wide and open digital library of syllabi.

However, there exists a chicken-and-egg situation regarding the adoption of such a repository on a much larger scale: there is little incentive for instructors to take the additional effort to add their syllabi to this repository unless there are existing services that they can then use. On the other hand, useful services would need a large collection of syllabi to work on. Hence, to break out of this deadlock, we decided to seed our repository with syllabi acquired from the Web in order to bootstrap the process. We restrict our focus to computer science syllabi offered by universities in the USA as a starting point of our proof-of-concept project. The methodology and the system could be extended easily to other disciplines and locations.

This paper presents our progress regarding automatic classification towards building a syllabus collection. A classification task usually can be accomplished by defining classes, selecting features, preparing a training corpus, and building a classifier. In order to build quickly an initial collection of CS syllabi, we obtained more than 8000 possible syllabus pages by automatically searching on Google. After randomly examining the set, we found the result set very noisy. To help with the task of properly identifying true syllabi, we defined four syllabus class types, shown in Table I, and then proposed syllabus feature characteristics for each class. We prepared a variety of training data in terms of their sizes and their distributions. Finally, we applied Support Vector Machines (SVMs) (Cortes & Vapnik, 1995) to learn classifiers to produce the syllabus repository.

There are many other genres of data on the Web. We hope that our application of machine learning techniques to obtain a repository of genre-specific data will encourage the creation of similar systems for other genres.

**CLASS DEFINITION**

The four classes of syllabi are defined in Table 1. A syllabus component is one of the following information: course code, title, class time and location, offering institute, teaching staffs, course description, objectives, Web site, prerequisite, textbook, grading policy, schedule, assignment, exam and resources. We consider only the full and the partial classes as syllabi. The reason we treat a partial syllabus as a syllabus that we can complete a partial syllabus by following outgoing links from it, which would be helpful for a variety of services. For example, in order to recommend papers or textbooks for a course using a partial syllabus, it is inaccurate just to extract frequent words from its syllabus since more features of the course are described in other pages. Therefore, we would like to recognize partial syllabi and then retrieve more complete information from them. Similarly, we also need to differentiate between an entry page and a noise page, although we consider neither of them as syllabi.
Related Content

A Workload Assignment Strategy for Efficient ROLAP Data Cube Computation in Distributed Systems
Ilhyun Suh and Yon Dohn Chung (2016). *International Journal of Data Warehousing and Mining* (pp. 51-71).
[www.igi-global.com/article/a-workload-assignment-strategy-for-efficient-rolap-data-cube-computation-in-distributed-systems/168486?camid=4v1a](www.igi-global.com/article/a-workload-assignment-strategy-for-efficient-rolap-data-cube-computation-in-distributed-systems/168486?camid=4v1a)

Measuring Semantic-Based Structural Similarity in Multi-Relational Networks
Yunchuan Sun, Rongfang Bie and Junsheng Zhang (2016). *International Journal of Data Warehousing and Mining* (pp. 20-33).

Data Mining for Human Resource Information Systems
[www.igi-global.com/chapter/data-mining-human-resource-information/7609?camid=4v1a](www.igi-global.com/chapter/data-mining-human-resource-information/7609?camid=4v1a)

Effectiveness of Fuzzy Classifier Rules in Capturing Correlations between Genes
[www.igi-global.com/article/effectiveness-fuzzy-classifier-rules-capturing/1818?camid=4v1a](www.igi-global.com/article/effectiveness-fuzzy-classifier-rules-capturing/1818?camid=4v1a)