Automatic Syllabus Classification

Xiaoyan Yu, Manas Tungare, Weiguo Fan, Manuel Perez-Quinones, and Edward A. Fox
Virginia Tech
Blacksburg, VA, USA
{xiaoyany, manas, wfan, perez, fox}@vt.edu

William Cameron, GuoFang Teng, and Lillian Cassel
Villanova University
Villanova, PA, USA
{william.cameron, guofang.teng, lillian.cassel}@villanova.edu

ABSTRACT

Syllabi are important educational resources. However, searching for a syllabus on the Web using a generic search engine is an error-prone process and often yields too many non-relevant links. In this paper, we present a syllabus classifier to filter noise out from search results. We discuss various steps in the classification process, including class definition, training data preparation, feature selection, and classifier building using SVM and Naïve Bayes. Empirical results indicate that the best version of our method achieves a high classification accuracy, i.e., an F_1 value of 83% on average.

Categories and Subject Descriptors: H.3 [Information Storage and Retrieval]: Digital Libraries

General Terms: Performance, Design, Experimentation

Keywords: Syllabus, Genre, Text Classification, SVM

1. INTRODUCTION

A course syllabus is the skeleton of a course. It is an important educational resource for educators, students, and life-long learners. 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 non-relevant links. The MIT OpenCourseWare[1] project, 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.

[1]http://ocw.mit.edu/

JCDL'07, June 18–23, 2007, Vancouver, British Columbia, Canada.

Class	Definition	Syllabus	Out-Links
Full	a syllabus without links to other syllabus components.	T	F
Partial	a syllabus along with links to other syllabus components somewhere else.	T	T
Entry Page	a page that contains a link to a syllabus.	F	T
Noise	all others.	F	N/A

Table 1: Class definitions.

This paper presents our progress regarding automatic classification towards a syllabus collection. A classification task usually can be accomplished by defining classes, selecting features, and building a classifier. In order to quickly build an initial collection of CS syllabi, we obtained more than 8000 possible syllabus pages by searching on Google (details in [4]). After randomly examining the set, we found the result set to be very noisy. To help with the task of properly identifying true syllabi, we defined four syllabus class types, shown in Table 1, and then proposed syllabus feature characteristics for each class. We used both content and structure information for syllabus classification, as they have been found useful in the detection of other genres [2]. Finally, we applied machine learning approaches to learn classifiers to produce the syllabus repository. We chose and compared Naïve Bayes and Support Vector Machine (SVM) approaches and their variances in the syllabus classification since they are good at text classification tasks in general [1, 3].

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.

2. CLASS DEFINITION AND FEATURE SELECTION

We define the four classes of documents acquired from the Web search in Table 1. We consider only the full and the partial classes as syllabi. The reason we treat a partial syllabus as a syllabus is that we can complete a partial syllabus by following outgoing links from it, which would be helpful for a variety of services. Each class has its own characteristics. An entry page contains a link that contains the word 'syllabus' or the prefix 'syl' or a link whose anchor text contains the syllabus keyword. Many keywords such as 'prerequisite' would occur in a full syllabus. These keywords would

F_1	NB-K	NB	SMO-L	SMO-P
Full	0.688	0.636	**0.752**	0.685
EntryPage	0.286	0.525	**0.629**	0.559
Partial	0.094	**0.386**	0.174	0.358
Noise	0.166	0.519	**0.597**	0.52
Avg	0.309	0.517	**0.538**	0.531
Acc_{tr}	56%	57%	71%	99%

Table 2: F_1 comparisons on Naïve Bayes (NB), Naïve Bayes with kernel (NB-K), SMO with linear kernel (SMO-L), and SMO with polynomial kernel of degree 5 and lambda=10 (SMO-P) on the original data set. Acc_{tr}: Accuracy on the training set.

F_1	NB-K	NB	SMO-L	SMO-P	INC_{PL}	INC_{43}
Full	0.569	0.514	0.607	**0.767**	26%	2%
Entry	0.568	0.51	0.73	**0.922**	26%	47%
Partial	0.551	0.477	0.581	**0.767**	32%	99%
Noise	0.67	0.564	0.747	**0.874**	17%	46%
Avg	0.590	0.516	0.666	**0.833**	25%	55%
Acc_{tr}	68%	54%	77%	99%	-	-

Table 3: F_1 comparisons on the same four settings on the data set with the uniform distribution after 100% re-sampling the original data set. Entry: EntryPage; Acc_{tr}: Accuracy on the training set; INC_{PL}: the increasing percentage that SMO-P is over SMO-L; INC_{43}: the increasing percentage that SMO-P is over the best results in the setting shown in bold in Table 2.

also occur in a partial syllabus but often along with a link. In addition, the position of a keyword in a page matters. For example, a keyword within the anchor text of a link or around the link would suggest a syllabus component outside the page. A capitalized keyword at the beginning of a page would suggest a syllabus component with a heading in the page. Motivated by the above observations, we selected 84 features mainly concerning the occurrences of keywords, the positions of keywords, and the co-occurrences of keywords and links.

3. EVALUATION

In order to train and evaluate our syllabus classifier, we randomly sampled 1020 documents from the web-acquired collection of more than 8000 documents and manually classified them into the four classes. We observed 499 full, 208 partial, 138 entry and 175 noise pages in the sample set. We extracted free text from the sampled documents in a variety of formats, along with their features, described in the above section, for the classification task, which was done with Naïve Bayes and SVM methods. We employed Naïve Bayes with/without kernel density estimation for numerical attributes and SMO (Sequential Minimal Optimization, a member of SVM family) with linear/polynomial kernel, all implemented in Weka[2].

Since data points for each class are not evenly distributed, we also use the re-sampling approach to create another set of data with each class containing evenly distributed data points. We trained and tested the classifiers on both data sets with 10-fold cross-validation and the average results are reported below.

F_1 values for the two data sets are shown in Tables 2 and Table 3. F_1 is a measure that balances precision and recall, to provide an overall measure of classification performance. For the data set with the original distribution, SMO with the linear kernel performs the best, but with F only 53.8% on average. For the data set after re-sampling, the performance of each classifier improves. Overall, the SMO with the polynomial kernel performs 25% better than the SMO with the linear kernel, and 55% better than the one in the original distribution setting in terms of average F_1. It also improves for each class, as can be seen in the last two columns in Table 3.

The training set accuracy for each setting is shown in Table 2 and 3. Generally SMO produces more accurate classifiers on the training set than Naïve Bayes. However, only SMO with the polynomial kernel can reach near 100% accuracy.

From these experiments, we come to the tentative conclusion that SVM performs better than Naïve Bayes for the syllabus classification task. SVM with a polynomial kernel performs the best. Our results also showed that the classifiers which are learnt from the data set with the uniform distribution perform better.

We did some failure analysis on the classification result. For our best result, SMO with the polynomial kernel on a uniform distribution data set, 61% of mis-labeled full syllabi were labeled as partial and 44% of the mis-labeled partial syllabi were labeled as full. In order to improve the classification between the full and the partial, we need to improve feature selection methods. For example, we considered the occurrences of the keywords around a link and among the anchor text of the link as the same feature. It would mislead the classifier to classify a full syllabus with links to a few resources as a partial one.

4. CONCLUSIONS

We presented a description of our approach to syllabus classification in this paper. We are still testing more settings, such as different feature selection approaches and training data set sizes, to discover more insights to build the best classifier for syllabi. We also are building an extractor for structured syllabi and creating services upon these collections.

5. ACKNOWLEDGMENTS

This work was funded in part by the National Science Foundation under DUE grant #0532825.

6. REFERENCES

[1] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning, Springer*, 1998.

[2] A. Kennedy and M. Shepherd. Automatic identification of home pages on the web. In *HICSS '05: Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 4*, Washington, DC, USA, 2005. IEEE Computer Society.

[3] S.-B. Kim, K.-S. Han, H.-C. Rim, and S. H. Myaeng. Some effective techniques for naive bayes text classification. *IEEE Transactions on Knowledge and Data Engineering*, 18(11):1457–1466, November 2006.

[4] M. Tungare, X. Yu, W. Cameron, G. Teng, M. Pérez-Quiñones, E. Fox, W. Fan, and L. Cassel. Towards a syllabus repository for computer science courses. In *Proceedings of the 38th Technical Symposium on Computer Science Education (SIGCSE 2007)*, 2007.

[2]http://www.cs.waikato.ac.nz/ml/weka/