Automatic Genre-Specific Text Classification

Xiaoyan Yu
Virginia Tech, USA

Manas Tungare
Virginia Tech, USA

Weiguo Fan
Virginia Tech, USA

Manuel Pérez-Quiñones
Virginia Tech, USA

Edward A. Fox
Virginia Tech, USA

William Cameron
Villanova University, USA

Lillian Cassel
Villanova University, USA

INTRODUCTION

Starting with a vast number of unstructured or semi-structured documents, text mining tools analyze and sift through them to present to users more valuable information specific to their information needs. The technologies in text mining include information extraction, topic tracking, summarization, categorization/classification, clustering, concept linkage, information visualization, and question answering [Fan, Wallace, Rich, & Zhang, 2006]. In this chapter, we share our hands-on experience with one specific text mining task — text classification [Sebastiani, 2002].

Information occurs in various formats, and some formats have a specific structure or specific information that they contain: we refer to these as *`genres'*. Examples of information genres include news items, reports, academic articles, etc. In this paper, we deal with a specific genre type, course syllabus.

A course syllabus is such a genre, with the following commonly-occurring fields: title, description, instructor's name, textbook details, class schedule, etc. In essence, a course syllabus is the skeleton of a course. Free and fast access to a collection of syllabi in a structured format could have a significant impact on education, especially for educators and life-long learners. Educators can borrow ideas from others' syllabi to organize their own classes. It also will be easy for life-long learners to find popular textbooks and even important chapters when they would like to learn a course on their own. Unfortunately, searching for a syllabus on the Web using Information Retrieval [Baeza-Yates & Ribeiro-Neto, 1999] techniques employed by a generic search engine often yields too many non-relevant search result pages (i.e., noise) — some of these only provide guidelines on syllabus creation; some only provide a schedule for a course event; some have outgoing links to syllabi (e.g. a course list page of an academic department). Therefore, a well-designed classifier for the search results is needed, that would help not only to filter noise out, but also to identify more relevant and useful syllabi.

This chapter presents our work regarding automatic recognition of syllabus pages through text classification to build a syllabus collection. Issues related to the selection of appropriate features as well as classifier model construction using both generative models (Naïve Bayes – NB [John & Langley, 1995; Kim, Han, Rim, & Myaeng, 2006]) and discriminative counterparts (Support Vector Machines – SVM [Boser, Guyon, & Vapnik, 1992]) are discussed. Our results show that SVM outperforms NB in recognizing true syllabi.

Copyright © 2009, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.

BACKGROUND

There has been recent interest in collecting and studying the syllabus genre. A small set of digital library course syllabi was manually collected and carefully analyzed, especially with respect to their reading lists, in order to define the digital library curriculum [Pomerantz, Oh, Yang, Fox, & Wildemuth, 2006]. In the MIT OpenCourseWare project, 1,400 MIT course syllabi were manually collected and made publicly available, which required a lot of work by students and faculty.

Some efforts have already been devoted to automating the syllabus collection process. A syllabus acquisition approach similar to ours is described in [Matsunaga, Yamada, Ito, & Hirokaw, 2003]. However, their work differs from ours in the way syllabi are identified. They crawled Web pages from Japanese universities and sifted through them using a thesaurus with common words which occur often in syllabi. A decision tree was used to classify syllabus pages and entry pages (for example, a page containing links to all the syllabi of a particular course over time). Similarly, [Thompson, Smarr, Nguyen, & Manning, 2003] used a classification approach to classify education resources – especially syllabi, assignments, exams, and tutorials. Using the word features of each document, the authors were able to achieve very good performance (F_1 score: 0.98). However, this result is based upon their relative clean data set, only including the four kinds of education resources, which still took efforts to collect. We, on the other hand, to better apply to a variety of data domains, test and report our approach on search results for syllabi on the Web.

In addition, our genre feature selection work is also inspired by research on genre classification, which aims to classify data according to genre types by selecting features that distinguish one genre from another, i.e., identifying home pages in sets of web pages [Kennedy & Shepherd, 2005].

MAIN FOCUS

A text 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 programmatically searching using Google [Tungare et al., 2007]. After randomly examining the result set, we found it to contain many documents that were not truly syllabi: we refer to this as noise. To help with the task of properly identifying true syllabi, we defined true syllabi and false syllabi, and then selected features specific to the syllabus genre. We randomly sampled the collection to prepare a training corpus of size 1020. All 1020 files were in one of the following formats: HTML, PDF, PostScript, or Text. Finally, we applied Naïve Bayes, Support Vector Machines, and its variants to learn classifiers to produce the syllabus repository.

Class Definition

A syllabus component is one of the following: course code, title, class time, class location, offering institute, teaching staff, course description, objectives, web site, prerequisite, textbook, grading policy, schedule, assignment, exam, or resource. A true syllabus is a page that describes a course by including most of these syllabus components, which can be located in the current page or be obtained by following outgoing links. A false syllabus (or noise) is a page for other purposes (such as an instructor's homepage with a link to syllabi for his/her teaching purpose) instead of describing a course.

The two class labels were assigned by three team members to the 1020 samples with unanimous agreement. A skewed class distribution was observed in the sample set with 707 true syllabus and 313 false syllabus pages. We used this sample set as our training corpus.

Feature Selection

In a text classification task, a document is represented as a vector of features usually from a high dimensional space that consists of unique words occurring in documents. A good feature selection method reduces the feature space so that most learning algorithms can handle and contribute to high classification accuracy. We applied three feature selection methods in our study: general feature selection, genre-specific feature selection, and a hybrid of the two.

1. *General Features* - In a study of feature selection methods for text categorization tasks [Yang & Pedersen, 1997], the authors concluded that Document Frequency (DF) is a good choice since

6 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/automatic-genre-specific-text-classification/10808

Related Content

Constraint-Based Association Rule Mining
Carson Kai-Sang Leung (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 307-312).*
www.irma-international.org/chapter/constraint-based-association-rule-mining/10837

Modeling Score Distributions
Anca Doloc-Mihu (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1330-1336).*
www.irma-international.org/chapter/modeling-score-distributions/10994

Behavioral Pattern-Based Customer Segmentation
Yinghui Yang (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 140-145).*
www.irma-international.org/chapter/behavioral-pattern-based-customer-segmentation/10811

Sampling Methods in Approximate Query Answering Systems
Gautam Das (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1702-1707).*
www.irma-international.org/chapter/sampling-methods-approximate-query-answering/11047

Vertical Data Mining on Very Large Data Sets
William Perrizo, Qiang Ding, Qin Dingand Taufik Abidin (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 2036-2041).*
www.irma-international.org/chapter/vertical-data-mining-very-large/11099