Chapter XXIII

Restrictive Methods and Meta Methods for Thematically Focused Web Exploration

Sergej Sizov
University of Koblenz-Landau, Germany

Stefan Siersdorfer
University of Sheffield, UK

ABSTRACT

This chapter addresses the problem of automatically organizing heterogeneous collections of Web documents for the generation of thematically-focused expert search engines and portals. As a possible application scenario for our techniques, we consider a focused Web crawler that aims to populate topics of interest by automatically categorizing newly-fetched documents. A higher accuracy of the underlying supervised (classification) and unsupervised (clustering) methods is achieved by leaving out uncertain documents rather than assigning them to inappropriate topics or clusters with low confidence. We introduce a formal probabilistic model for ensemble-based meta methods and explain how it can be used for constructing estimators and for quality-oriented tuning. Furthermore, we provide a comprehensive experimental study of the proposed meta methodology and realistic use-case examples.

INTRODUCTION

The novel Web exploration paradigm of focused crawling or topical crawling can be viewed as an attempt to automate intellectual pre-processing and post-processing of Web information (Chakrabarti, 2003; Sizov, Biwer, Graumann, Siersdorfer, Theobald, Weikum, & Zimmer, 2003; Sizov, Siersdorfer, Theobald, & Weikum, 2002). In contrast to a search engine’s generic crawler (which serves to build and maintain the engine’s index), a focused crawler is interested only in a specific, typically small, set of topics. Each of the visited documents is classified into the crawler’s hierarchy of topics to test whether it is of interest at all and where it belongs within a user-specific taxonomy. This step is automated using classification techniques from machine learning such as Naive Bayes, Maximum Entropy, Support Vector Machines (SVM), or other supervised learning...
Restrictive Methods and Meta Methods for Thematically Focused Web Exploration

methods. The outcome of the focused crawl can be viewed as the index of a personalized information service or a thematically-specialized search engine. The crawler, along with its document repository, resides on the user’s local computer and does not require any centralized services, making the approach robust, objective, and privacy-preserving.

An important aspect of thematically-focused Web exploration applications is the recognition and elimination of thematically-irrelevant documents. Common IR methods usually address “closed” scenarios with a limited number of predefined categories. This is a very significant difference to the “open” Web scenario for which comprehensive learning of all existing topics and themes is clearly impossible. Ideally, a quality-oriented Web IR method should automatically recognize (and reject) documents that do not belong to the desired topics of interest. Another example of a post-processing task is the filtering of the data repository. As a result of the crawl evaluation, the user may decide to remove from the repository relevant but “uncertain” documents with low classification confidence. The reduced repository is expected to contain smaller but much more concise collections of highly-relevant documents.

In this chapter, we discuss meta methods based on ensembles of classification or clustering methods for quality-oriented organization of document collections. Our main focus lies on restrictive solutions: organizing only a subset of the data, but doing so with much higher accuracy. Our goal is to allow the user managing the crawl results in a robust way and with minimized human efforts by:

- elimination of thematically-irrelevant "junk" documents;
- robust and tunable categorization of crawl inputs;
- restrictive filtering of crawl results; and
- collaborative organization and filtering of relevant topics.

More generally, this chapter introduces widely-applicable techniques for improving the quality of Web systems. The methodology addressed here can be used for the largely-automated generation of thematically-focused portals and Web taxonomies, "needle-in-a-haystack" expert Web search, design of quality-oriented intelligent user interfaces, and other related applications.

RELATED WORK

There is a plethora of work on text document classification using a variety of probabilistic and discriminative models (Chakrabarti, 2002). The emphasis of this body of work has been on the mathematical and algorithmic aspects, and the engineering aspects of how to cope with trade-offs and how to tune a classifier with regard to properties of the training data.

The machine-learning literature has studied a variety of ensemble-based meta methods such as bagging, stacking, or boosting (Breiman, 1996; Freund, 1999; Kuncheva, 2004; Littlestone & Warmuth, 1989; Wolpert, 1992) and also combinations of heterogeneous learners (e.g., Yu, Chang, & Han, 2002). For bagging, an ensemble consists of classifiers built on bootstrap replicates of the training set. The classifiers outputs are combined by the plurality vote. For stacking, multiple classifiers are trained on parts of the training set and evaluated on the remaining training documents. The outputs of the classifiers are used as feature values for training a new classifier (stacked generalization). Boosting can be viewed as a model averaging method. Here a succession of models is built, each one trained on a data set in which the points misclassified by the previous model are given more weight.

The approach of intentionally splitting training sets for meta learning has been investigated by Chan (1996). The solution proposed in Schein, Popescul, Ungar, and Pennock (2002) studied the accuracy-loss trade-off in a ROC curve model (for a recommender system).
Related Content

Web Technology Systems Integration Using SOA and Web Services
[www.igi-global.com/chapter/web-technology-systems-integration-using-soa-and-web-services/203415?camid=4v1a](www.igi-global.com/chapter/web-technology-systems-integration-using-soa-and-web-services/203415?camid=4v1a)

Successful Web-Based IT Support Services: Service Provider Perceptions of Stakeholder-Oriented Challenges
[www.igi-global.com/chapter/successful-web-based-support-services/37677?camid=4v1a](www.igi-global.com/chapter/successful-web-based-support-services/37677?camid=4v1a)

A Novel Approach to Design a Scalable Air Quality Monitoring System
[www.igi-global.com/article/novel-approach-design-scalable-air/72991?camid=4v1a](www.igi-global.com/article/novel-approach-design-scalable-air/72991?camid=4v1a)

A Study on Components and Features in Face Detection
[www.igi-global.com/article/a-study-on-components-and-features-in-face-detection/145839?camid=4v1a](www.igi-global.com/article/a-study-on-components-and-features-in-face-detection/145839?camid=4v1a)