Book Review

Algorithms of the Intelligent Web

Haralambos Marmanis and Dmitry Babenko
Manning Publications Co., 2009
ISBN 978-1-933988-66-5

"Algorithms of the Intelligent Web" is a timely and instructive book that brings together techniques for the rapidly evolving field of intelligent web applications. It covers problems such as crawling, indexing, searching, integrating search results with link analysis and user's clicks, generating recommendations, clustering and classification. The reader will become familiar with a large number of datasets, open-source libraries and APIs for developing intelligent web-based applications. The example-driven approach used by the authors puts the reader on a fast track to understanding the major techniques used in current applications. The emphasis is on practice rather than theory. Most of the techniques are presented in the form of algorithms without placing heavy emphasis on their mathematical aspects.

The book consists of seven chapters and five appendices. Chapter 1 is a general introduction to the topic of intelligent web. It briefly reviews the history of intelligent web applications with a number of examples based on real websites, and summarizes the elements that are usually required in order to build these applications. It presents typical scenarios where the algorithms introduced in the rest of the book can be utilized, such as social networking, mashups, portals, wikis, media-sharing sites and online gaming. This introductory chapter also gives an overview of the relation between intelligent web applications and the fields of artificial intelligence, machine learning, data mining and soft computing. Finally, it indicates and discusses the most common difficulties that are found in practice.

Chapter 2 summarizes the features of the Lucene library, a high-performance, full-featured text search engine library written in the Java programming language. It highlights and motivates the importance of improving search results with link analysis and introduces the PageRank algorithm. The discussion of the PageRank algorithm is done in a simple yet very informative manner. It shows how to build the hyperlink matrix and the PageRank vector, pointing to important practical aspects such as the stochasticity adjustment and the primitivity adjustment of the hyperlink matrix. It then shows how to integrate the Lucene search scores with the PageRank scores. This chapter also brings web usage mining into play by discussing how to take advantage of user clicks to improve search results. It then presents DocRank, an algorithm proposed by the authors for ranking Word, PDF and other documents without links. It continues with a discussion of large-scale implementation issues, pointing to useful techniques and tools that can help address memory and speed constraints. The chapter concludes with a review of the precision and recall evaluation metrics.

Chapter 3 is a pragmatic introduction to the topic of recommendation systems. It begins by discussing basic concepts such as distance and similarity between users and items. This is followed by a review of the two main categories of recommendation engines: collaborative filtering and the content-based approach. The chapter then moves on to illustrate the concepts presented so far by showing how to build realistic examples of recommendation engines. These examples include a system for recommending friends, articles and news stories as well as a system for recommending movies. Useful pointers to datasets and programming resources are indicated for both examples. The chapter also presents some helpful guidelines for evaluating recommendation systems and then closes with a discussion of large-scale implementation and evaluation issues.

Chapter 4 is a general overview of clustering algorithms. It first covers some simple approaches to clustering and then progresses to more advanced mathematical techniques. Six clustering algorithms are discussed in detail: single link, average link, minimum-spanning tree, k-means, ROCK, and DBSCAN. The presentation of these algorithms allows introducing several fundamental concepts such as dendograms, centroids, density, chain effect, noise, and outliers, among others. General clustering issues as well as advantages and disadvantages of each algorithm are discussed, including an analysis of their space and time complexity.

Chapter 5 introduces the topic of classification and starts by providing a number of real-world examples where classification is important. It then presents a concise overview of structural and statistical classifiers and a general description of the typical lifecycle of a classifier. The chapter continues with a closer look to specific

classifiers, which is accompanied with concrete classification problems. In particular, naïve-based classifiers are used for the task of filtering spam messages and for placing email messages in several appropriate folders. In order to illustrate the application of a different classifier on the same domain, the Drool rule language is introduced and used to implement a rule-based classifier that identifies spam email. The chapter then presents detailed guidelines for the implementation of a fraud detector neural network. Finally, the chapter describes the most common tests used to evaluate classifiers and provides tips for dealing with very large datasets.

Chapter 6 presents the topic of combining multiple classifiers. The practical relevance of this topic is illustrated by using as a case study the evaluation of the credit worthiness of mortgage applicants. After presenting a number of tests for comparing multiple classifiers the rest of the chapter focuses on how to combine them using classifier fusion, a general approach where all classifiers contribute to a given classification. Two specific techniques to ensemble classification are discussed in detail: bagging and boosting. Special attention is given to the implementation, extension and evaluation of these two techniques.

Chapter 7 shows how to adopt the studied techniques in the implementation of a news portal. The chapter reviews the task of collecting and cleansing content, indexing and searching news stories, clustering news stories, classifying them into topics and generating recommendations based on the user's ratings. In other words, this chapter puts it all together in the context of an intelligent web application.

The five appendices provide supporting material on selected tools and topics. This includes the BeanShell scripting language, an overview of crawler components, a mathematical refresher that briefly reviews the concepts of vectors and matrices, some pointers to natural language processing resources and references to literature on neural networks.

An important part of this book are the To Do sections at the end of each chapter. These sections propose exercises that entice the reader to go deeper into the topics covered in the chapters. The to-do items serve as a starting point for discussion and potential exploration of novel ideas and tools. Another main component of this book is the code, which can be downloaded from the book website and is carefully explained as it is introduced.

Dr. Haralambos Marmanis and Dmitry Babenko, the authors of this book, are experienced practitioners in the domain of business intelligence and in the adoption of machine learning techniques for industrial solutions. According to them, this book is intended for software professionals, but I personally believe that it is also highly recommendable for students. It may not be suitable as the main course book to introduce the most advanced mathematical aspects of web mining and information retrieval, but it is definitely an excellent resource to help understand and build real-world web applications.

Ana Gabriela Maguitman
agm@cs.uns.edu.ar