Abstract. A very common issue of adaptive Web-Based systems is the modeling of documents. Such documents represent domain-specific information for a number of purposes. Application areas such as Information Search, Focused Crawling and Content Adaptation (among many others) benefit from several techniques and approaches to model documents effectively. For example, a document usually needs preliminary processing in order to obtain the relevant information in an effective and useful format, so as to be automatically processed by the system. The objective of this chapter is to support other chapters, providing a basic overview of the most common and useful techniques and approaches related with document modeling. This chapter describes high-level techniques to model Web documents, such as the Vector Space Model and a number of AI approaches, such as Semantic Networks, Neural Networks and Bayesian Networks. This chapter is not meant to act as a substitute of more comprehensive discussions about the topics presented. Rather, it provides a brief and informal introduction to the main concepts of document modeling, also focusing on the systems that are presented in the rest of the book as concrete examples of the related concepts.

5.1 Introduction

The Web document, in its various representation forms, is the focal point of this chapter, which aims at illustrating the most common techniques employed in Web Document Modeling literature, with particular attention given to those used in Adaptive Web-Based systems. The purpose of this chapter is to offer a support for the other chapters in this volume dealing with the various adaptive systems in greater detail. The present chapter is therefore not to be considered a comprehensive guide to the discussed topics, nor a substitute for more specialized literature. Readers are encouraged to consult the provided references in order to broaden their grasp of the discussed topics or equivalent literature.

1 In this context, with the term modeling we mean the construction of an abstract representation of the document, useful for all applications aimed at processing information automatically.

2 Fig. 5.1 shows the structure of this chapter together with the links to the other chapters of this book.
It is well known that, owing to the Internet, each one of us can benefit from a large quantity of information, available on-line in several, essentially standard, formats, such as HTML, XML and XHTML text pages and jpeg and tif graphic ones. More complex formats, particularly multimedia (audio and/or video) usually require a longer search and interaction with the Web, as well as more sophisticated fruition tools installed on user clients, now widely accessible.

The quantity of information available on the World Wide Web is increasing exponentially, and this boom has paved the way for a new era, creating new opportunities in many different fields, such as e-business, e-commerce, e-marketing, e-finance and e-learning, just to mention some of the most interesting ones. Web Intelligence [90, 91] and Wisdom Web [93, 48, 92] are other examples of new disciplines born from the development of the World Wide Web. All this proceeds from the birth of HTTP, HyperText Transfer Protocol [80, 39] and of HTML, HyperText Markup Language, a subset of SGML, Standard Generalized Markup Language [29], which made the Internet enjoyable for anyone who had a computer with a browser on-board. With time, the number of Web surfers increased, and so did the available Web documents, especially HTML pages. This eventually required the need to gather the information to be supplied in more structured containers, enabling more thorough, personalized searches, directly correlated to the semantics of the document itself, for a more intelligent fruition and to provide more sophisticated search-tools systems.

All the aforementioned techniques are the ones used concretely to represent documents on various Web sites, and may therefore be called Layout Representations. For automatic systems that operate on the Web, however, e.g. adaptive Web-Based systems, such documents often need to be pre-processed, that is, recasting them in representations different from primitive ones and more apt to undergo a particular elaboration. Think of a personalized search system in which the calculation of the relevance of a

---

3 The reader can visit the w3c Web site: http://www.w3c.org for further reading on Web standards and protocols.