With the continued growth and proliferation of e-commerce, Web services, and Web-based information systems, the volumes of clickstream, transaction data, and user profile data collected by Web-based organizations in their daily operations has reached astronomical proportions. Analyzing such data can help these organizations determine the life-time value of clients, design cross-marketing strategies across products and services, evaluate the effectiveness of promotional campaigns, optimize the functionality of Web-based applications, provide more personalized content to visitors, and find the most effective logical structure for their Web space. This type of analysis involves the automatic discovery of meaningful patterns and relationships from a large collection of primarily semi-structured data, often stored in Web and applications server access logs, as well as in related operational data sources.

Web usage mining refers to the automatic discovery and analysis of patterns in clickstreams, user transactions and other associated data collected or generated as a result of user interactions with Web resources on one or more Web sites [28, 82, 118]. The goal is to capture, model, and analyze the behavioral patterns and profiles of users interacting with a Web site. The discovered patterns are usually represented as collections of pages, objects, or resources that are frequently accessed or used by groups of users with common needs or interests.

Following the standard data mining process [38], the overall Web usage mining process can be divided into three inter-dependent stages: data collection and pre-processing, pattern discovery, and pattern analysis. In the pre-processing stage, the clickstream data is cleaned and partitioned into a set of user transactions representing the activities of each user during different visits to the site. Other sources of knowledge such as the site content or structure, as well as semantic domain knowledge from site ontologies (such as product catalogs or concept hierarchies), may also be used in pre-processing or to enhance user transaction data. In the pattern discovery stage, statistical, database, and machine learning operations are performed to obtain hidden patterns reflecting the typical behavior of users, as well as summary statistics on Web resources, sessions, and users. In the final stage of the process, the discovered patterns and statistics are...
further processed, filtered, possibly resulting in aggregate user models that can be used as input to applications such as recommendation engines, visualization tools, and Web analytics and report generation tools. The overall process is depicted in Fig. 12.1.

In the remainder of this chapter, we first provide a detailed examination of Web usage mining as a process, and discuss the relevant concepts and techniques commonly used in all the stages mentioned above. We then focus on the important problem of recommendation. It is followed by another special case of Web usage mining targeting search query logs, and known as query log mining (QLM). To complete our discussion of mining Web usage data, we finally introduce the new field of Ad click mining.

12.1 Data Collection and Pre-Processing

An important task in any data mining application is the creation of a suitable target data set to which data mining and statistical algorithms can be applied. This is particularly important in Web usage mining due to the characteristics of clickstream data and its relationship to other related data collected from multiple sources and across multiple channels. The data