Web mining research support system computer science essay

Technology, Computer

Application of data mining techniques to the World Wide Web, referred to as Web mining has been the focus of several recent research projects and papers. However, there is no established vocabulary leading to confusion when comparing research efforts. The term Web mining has been used in two distinct ways. The first, called Web content mining in this paper is the process of information discovery from sources across the World Wide Web. The second called Web usage mining is the process of mining for user browsing and access patterns. In this paper we define Web mining and present an overview of the various research issues, techniques, and development efforts. We briefly describe WEBMINER, a system for Web usage mining, and conclude this paper by listing research issues.

Introduction

The evolution of the World Wide Web has brought us enormous and ever growing amounts of data and information. It inuences almost all aspects of people’s lives. In addition, with the abundant data provided by the web, it has become an important resource for research. Furthermore, the low cost of web data makes it more attractive to researchers.

Researchers can retrieve web data by browsing and keyword searching [58]. However, there are several limitations to these techniques. It is hard for researchers to retrieve data by browsing because there are many following links contained in a web page. Keyword searching will return a large amount of irrelevant data. On the other hand, traditional data extraction and mining techniques cannot be applied directly to the web due to its semi-structured or even unstructured nature. Web pages are Hypertext documents, which

contain both text and hyperlinks to other documents. Furthermore, other data sources also exist, such as mailing lists, newsgroups, forums, etc. Thus, design and implementation of a web mining research support system has become a challenge for people with interest in utilizing information from the web for their research.

A web mining research support system should be able to identify web sources according to research needs, including identifying availability, relevance and importance of web sites; it should be able to select data to be extracted, because a web site 1 contains both relevant and irrelevant information; it should be able to analyze the data patterns of the collected data and help to build models and provide validity.

A Taxonomy of Web Mining

In this section I will present the taxonomy of web mining. In fact, web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined. I will provide a description and categorization of some of the recent work. in addition to this, I will give some tools and techniques related to each area.

2-1. Web Content Mining

The lack of structure that permeates the information sources on the World Wide Web makes automated discovery of Web-based information difficult. Traditional search engines such as Lycos, Alta Vista WebCrawler, ALIWEB [29], MetaCrawler, and others provide some comfort to users, but do not

generally provide structural information nor categorize, filter, or interpret documents. A recent study provides a comprehensive and statistically thorough comparative evaluation of the most popular search engines.

In recent years these factors have pushed researchers to develop more intelligent tools for information retrieval such as intelligent Web agents, and to extend data mining techniques to provide a higher level of organization for semi-structured data available on the Web. We summarize some of these efforts below.

2-1-1. Agent-Based Approach:

Generally, agent-based Web mining systems can be placed into the following three categories.

Intelligent Search Agents: Several intelligent Web agents have been developed that search for relevant information using domain characteristics and user profiles to organize and interpret the discovered information. Agents such as Harvest [9] FAQ-Finder [19] Information Manifold [27] OCCAM [30] and ParaSite [51] rely either on pre-specified domain information about particular types of documents, or on hard coded models of the information sources to retrieve and interpret documents_ Agents such as Shop-Bot [14] and ILA (Internet Learning Agent) [42] interact with and learn the structure of unfamiliar information sources. ShopBot retrieves product information from a variety of vendor sites using only general information about the product domain. ILA learns models of various information sources and translates these into its own concept hierarchy.

Information Filtering/ Categorization: A number of Web agents use various information retrieval techniques [17] and characteristics of open hypertext Web documents to automatically retrieve, filter, and categorize them[5, 9, 34, 55, 53] HyPursuit [53] uses semantic information embedded in link structures and document content to create cluster hierarchies of hypertext documents, and structure an information space. BO (Bookmark Organizer [34] combines hierarchical clustering techniques and user interaction to organize a collection of Web documents based on conceptual information.

Personalized Web Agents: This category of Web agents learn user preferences and discover Web information sources based on these preferences, and those of other individuals with similar interests (using collaborative filtering). A few recent examples of such agents include the WebWatcher[3], PAINT [39]. Syskill & Webert [41]. GroupLens [47]. Firefly [49] and others [4]. For example, Syskill & Webert utilizes a user profile and learns to rate Web pages of interest using a Bayesian classifier.

2-1-2. Database Approach

Database approaches to Web mining have focused on techniques for organizing the semi-structured data on the Web into more structured collections of resources, and using standard database querying mechanisms and data mining techniques to analyze it.

Multilevel Databases:

The main idea behind this approach is that the lowest level of the database contains semi-structured information stored in various Web repositories, such as hypertext documents. At the higher level(s) meta data or generalizations are extracted from lower levels and organized in structured collections, i. e. relational or object oriented databases. For example, Han, et. al. use a multilayered database where each layer is obtained via generalization and transformation operations performed on the lower layers. Kholsa, et. al. propose the creation and maintenance of meta-databases at each information providing domain and the use of a global schema for the meta-database. King & Novak propose the incremental integration of a portion of the schema from each information source, rather than relying on a global heterogeneous database schema. The

ARANEUS system extracts relevant information from hypertext documents and integrates these into higher-level derived Web Hypertexts which are generalizations of the notion of database views.

Web Query Systems

Many Web-base query systems and languages utilize standard database query languages such as SQL, structural information about Web documents, and even natural language processing for accommodating the types of queries that are used in World Wide Web searches. We mention a few examples of these Web-base query systems here. W3Q combines structure queries, based on the organization of hypertext documents, and content queries, based on information retrieval techniques. WebLog Logic-based

query language for restructuring extracted information from Web information sources. Lorel and UnQL query heterogeneous and semi-structured information on the Web using a labeled graph data model. TSIMMIS extracts data from heterogeneous and semi-structured information sources and correlates them to generate an integrated database representation of the extracted information.