Web Data Mining and the Development of Knowledge-Based Decision Support Systems

G. Sreedhar

IGI GLOBAL
DISSEMINATOR OF KNOWLEDGE

Web Data Mining and the Development of Knowledge-Based Decision Support Systems

G. Sreedhar
Rashtriya Sanskrit Vidyapeetha (Deemed University), India

A volume in the Advances in Data Mining and Database Management (ADMDM) Book Series

www.igi-global.com

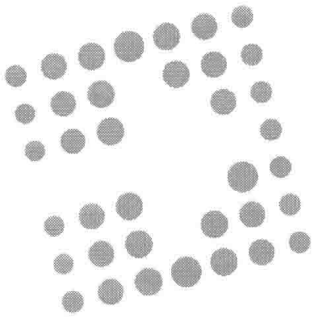
Published in the United States of America by
IGI Global
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA, USA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

CIP Data Pending
ISBN: 978-1-5225-1877-8
eISBN: 978-1-5225-1878-5

This book is published in the IGI Global book series Advances in Data Mining and Database Management (ADMDM) (ISSN: 2327-1981; eISSN: 2327-199X)

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.



Advances in Data Mining and Database Management (ADMDM) Book Series

David Taniar
Monash University, Australia

ISSN:2327-1981
EISSN:2327-199X

Mission

With the large amounts of information available to organizations in today's digital world, there is a need for continual research surrounding emerging methods and tools for collecting, analyzing, and storing data.

The **Advances in Data Mining & Database Management (ADMDM)** series aims to bring together research in information retrieval, data analysis, data warehousing, and related areas in order to become an ideal resource for those working and studying in these fields. IT professionals, software engineers, academicians and upper-level students will find titles within the ADMDM book series particularly useful for staying up-to-date on emerging research, theories, and applications in the fields of data mining and database management.

Coverage

- Customer Analytics
- Profiling Practices
- Educational Data Mining
- Web-based information systems
- Data Warehousing
- Data Analysis
- Database Security
- Web mining
- Text Mining
- Data Mining

IGI Global is currently accepting manuscripts for publication within this series. To submit a proposal for a volume in this series, please contact our Acquisition Editors at Acquisitions@igi-global.com or visit: http://www.igi-global.com/publish/.

The Advances in Data Mining and Database Management (ADMDM) Book Series (ISSN 2327-1981) is published by IGI Global, 701 E. Chocolate Avenue, Hershey, PA 17033-1240, USA, www.igi-global.com. This series is composed of titles available for purchase individually; each title is edited to be contextually exclusive from any other title within the series. For pricing and ordering information please visit http://www.igi-global.com/book-series/advances-data-mining-database-management/37146. Postmaster: Send all address changes to above address.

Titles in this Series

For a list of additional titles in this series, please visit: www.igi-global.com

Emerging Trends in the Development and Application of Composite Indicators
Veljko Jeremic (University of Belgrade, Serbia) Zoran Radojicic (University of Belgrade, Serbia) and Marina Dobrota (University of Belgrade, Serbia)
Information Science Reference • copyright 2017 • 402pp • H/C (ISBN: 9781522507147) • US $205.00 (our price)

Web Usage Mining Techniques and Applications Across Industries
A.V. Senthil Kumar (Hindusthan College of Arts and Science, India)
Information Science Reference • copyright 2017 • 424pp • H/C (ISBN: 9781522506133) • US $200.00 (our price)

Social Media Data Extraction and Content Analysis
Shalin Hai-Jew (Kansas State University, USA)
Information Science Reference • copyright 2017 • 493pp • H/C (ISBN: 9781522506485) • US $225.00 (our price)

Collaborative Filtering Using Data Mining and Analysis
Vishal Bhatnagar (Ambedkar Institute of Advanced Communication Technologies and Research, India)
Information Science Reference • copyright 2017 • 309pp • H/C (ISBN: 9781522504894) • US $195.00 (our price)

Effective Big Data Management and Opportunities for Implementation
Manoj Kumar Singh (Adama Science and Technology University, Ethiopia) and Dileep Kumar G. (Adama Science and Technology University, Ethiopia)
Information Science Reference • copyright 2016 • 324pp • H/C (ISBN: 9781522501824) • US $195.00 (our price)

Data Mining Trends and Applications in Criminal Science and Investigations
Omowunmi E. Isafiade (University of Cape Town, South Africa) and Antoine B. Bagula (University of the Western Cape, South Africa)
Information Science Reference • copyright 2016 • 386pp • H/C (ISBN: 9781522504634) • US $210.00 (our price)

Intelligent Techniques for Data Analysis in Diverse Settings
Numan Celebi (Sakarya University, Turkey)
Information Science Reference • copyright 2016 • 353pp • H/C (ISBN: 9781522500759) • US $195.00 (our price)

Managing and Processing Big Data in Cloud Computing
Rajkumar Kannan (King Faisal University, Saudi Arabia) Raihan Ur Rasool (King Faisal University, Saudi Arabia) Hai Jin (Huazhong University of Science and Technology, China) and S.R. Balasundaram (National Institute of Technology, Tiruchirappalli, India)
Information Science Reference • copyright 2016 • 307pp • H/C (ISBN: 9781466697676) • US $200.00 (our price)

www.igi-global.com

701 E. Chocolate Ave., Hershey, PA 17033
Order online at www.igi-global.com or call 717-533-8845 x100
To place a standing order for titles released in this series, contact: cust@igi-global.com
Mon-Fri 8:00 am - 5:00 pm (est) or fax 24 hours a day 717-533-8661

Acknowledgment

I am very much happy and thankful to IGI global for giving me the opportunity to produce my second *Web Data Mining and the Development of Knowledge-Based Decision Support Systems*, which is very much necessary in the present internet world. I express a deep sense of gratitude to the Jan Travers, David J Perreault, Eleana Wehr, Joshua Herring, Kayla Wolfe, and other members of IGI global who supported either directly or indirectly during book project development.

I am thankful to all authors who contributed their valuable efforts and ideas in the form of chapters in the book. I would like to express my sincere thanks to Prof. A. Anandaraja Chari for his continuous support and encouragement in my career.

I am thankful to vice chancellor, registrar and other faculty members of Rashtriya Sanskrit Vidyapeetha, Tirupati for their support and kindness in completing the book.

Finally, I am very much thankful to my parents and family members for their support and encouragement in achieving this target goal in my academic career.

Section 1

Developing Decision Support Systems

Table of Contents

Detailed Table of Contents

been a common practice to design online collaborative learning activities to enhance learning. To apply learning analytics techniques to monitor the online collaborative process enables a lecturer to make instant and informed pedagogical decisions. However, it is still a challenge to build strong connection between learning analytics and learning science for understanding cognitive progression in learning. In this connection, this chapter reports a study to apply learning analytics techniques in the aspect of web usage mining and clustering analysis with underpinning Bloom's taxonomy to analyze students' performance in the online collaborative learning process. The impacts of intermediate interventions are also elaborated.

The Web can be defined as a depot of varied range of information present in the form of millions of websites dispersed around us. Often users find it difficult to locate the appropriate information fulfilling their needs with the abundant number of websites in the Web. Hence multiple research work has been conducted in the field of Web Mining so as to present any information matching the user's needs. The application of data mining techniques on web usage, web content or web structure data to find out useful data like users' way in patterns and website utility statistics on a whole can be defined as Web mining. The main cause behind development of such websites was to personalize the substance of a website on user's preference. New methods are developed to deal with a Web site using a link hierarchy and a conceptual link hierarchy respectively on the basis of how users have used the Web site link structure.

In recent days, Internet technology has provided a lot of services for sharing and distributing information across the world. Among all the services, World Wide Web (WWW) plays a significant role. The slow retrieval of Web pages may lessen the interest of users from accessing them. To deal with this problem, Web caching and Web pre-fetching are the two techniques used. Web proxy caching plays a key role in improving Web performance by keeping Web objects that are likely to be used in the near future in the proxy server which is closer to the end user. It helps in reducing user perceived latency, network bandwidth utilization, and alleviating loads on the Web servers. Thus, it improves the efficiency and scalability of Web based system. This chapter gives an overview of Web usage mining and its application on Web and discusses various approaches for improving the performance of Web.

This chapter explains the multi-agent system for effective information retrieval using information scent in query log mining. The precision of search results is low due to difficult to infer the information need

of the small size search query and therefore information need of the user is not satisfied effectively. Information Scent is used for modeling the information need of user web search session and clustering is performed to identify the similar information need sessions. Hyper Link-Induced Topic Search (HITS) is executed on clusters to generate the Hubs and authorities for web page recommendations to users who search with similar intents. This multi-agent system based on clustered query sessions uses query operations like expansion and recommendation to infer the information need of user search queries and recommends Hubs and authorities for effective web search.

Section 3
Developing Efficient Knowledge based Systems

The World Wide Web (WWW) is global information medium, where users can read and write using computers over internet. Web is one of the services available on internet. The Web was created in 1989 by Sir Tim Berners-Lee. Since then a great refinement has done in the web usage and development of its applications. Semantic Web Technologies enable machines to interpret data published in a machine-interpretable form on the web. Semantic web is not a separate web it is an extension to the current web with additional semantics. Semantic technologies play a crucial role to provide data understandable to machines. To achieve machine understandable, we should add semantics to existing websites. With additional semantics, we can achieve next level web where knowledge repositories are available for better understanding of web data. This facilitates better search, accurate filtering and intelligent retrieval of data. This paper discusses about the Semantic Web and languages involved in describing documents in machine understandable format.

As the speed of information growth exceeds in this new century, excessive data is making great troubles to human beings. However, there are so much potential and highly useful values hidden in the huge volume of data. Big Data has drawn huge attention from researchers in information sciences, policy and decision makers in governments and enterprises. Data analytic is the science of examining raw data with the purpose of drawing conclusions about that information. Data analytics is about discovering knowledge from large volumes data and applying it to the business. Machine learning is ideal for exploiting the opportunities hidden in big data. This chapter able to discover and display the patterns buried in the data using machine learning.

Data mining extracts novel and useful knowledge from large repositories of data and has become an effective analysis and decision means in any organization. The resource of the World Wide Web is almost infinite. The growing importance of electronic media for storing and disseminating text documents has

created an urgent need for tools and techniques that assist users in finding and extracting relevant and previously unknown information from massive collection of documents available in the web. Thus the development of techniques for mining unstructured, semi-structured, and fully structured textual data has become quite important in both academia and industry. Information management of well organized databases has been a focus of the Data mining research. When to specify too many attributes, system will slow down thus exclude irrelevant or weakly relevant attributes. The general idea behind attribute relevance analysis is to compute some measure that is used to quantify the relevance of an attribute with respect to a given class or concept.

Chapter 12

The hotlinks are the special links introduced in the website to reduce the time to access certain webpages in a webpage that is present in the deeper levels of the topology. Hotlinks selection mechanism plays a vital role in quick access of webpages. The problem is to decide which webpage should be having hotlinks and where the hotlinks should be placed in the website tree topology. We have proposed a methodology which starts by finding the frequent webpage access pattern of visitors of the website. The frequent pattern is found using Associative mining, Apriori algorithm or Frequent Pattern Tree algorithm. Then the frequent patterns are passed through page ranking mechanism. We find the pattern which is having the highest priority. Then the hotlinks are created for the members (webpages hyperlinks) of the pattern. Thus, the work is about assigning hotlinks for a set of pages which are frequently visited. Thus, by updating the topology by introducing hotlinks we can reduce the time to access the web pages.

Section 4
Developing Social Media based Mining Systems

Chapter 13

Social media mining is the process of representing, analyzing, and extracting actionable patterns and trends from raw social media data. Social media is favored by many users since it is available to individuals without any limitations to share their opinions, educational learning experiences and concerns via their status. Twitter API, twitter4j, is processed for searching the tweets based on the geo location. Student's posts on social network offers us a stronger concern to take decisions concerning the particular education system's learning method of the system. Evaluating knowledge in social media is sort of a difficult method. Bayes classifier are enforced on deep-mined knowledge for analysis purpose to urge the deeper understanding of the information. It uses multi label classification technique as every label falls into completely different classes. Label based measures are mostly taken to research the results and comparing them with the prevailing sentiment analysis technique.