CHAPTER: 11

CONCEPTS OF WEB MINING

[1]Dr. SUDHAKAR RANJAN
[1]School of engineering and technology
Apeejay stya university Sohna-Palwal road, Gurugram

[2]Dr. PARIKSHIT VASISHT
[2]School of engineering and technology
Apeejay stya university Sohna-Palwal road, Gurugram

Ch.Id:-ASU/GRF/EB/RPETHEAT/2022/Ch-11
DOI: https://doi.org/10.52458/9789391842888.2022.eb.grf.asu.ch-11

INTRODUCTION

Determining the size of the WWW is very challenging. The following Sources of data for Web mining can be collected at one of these three parts.

- Server level collection
- Client level collection
- Proxy level collection

The web data can be classified into the following modules.

- Content of web pages
- Intra page structure
- HTML or XML code for the pages
- Web pages accessed by the visitors
- User Profile

The access pattern record in server log shown below:

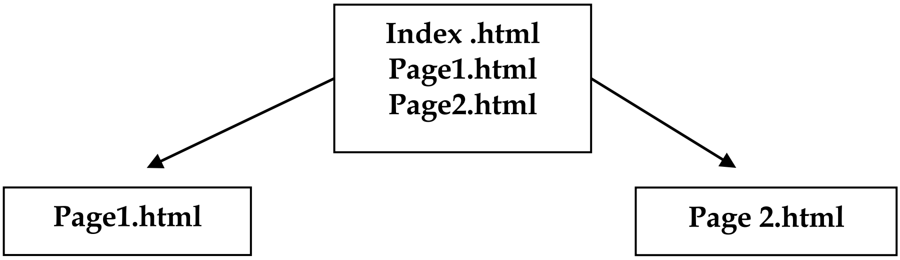

Fig. 1 Access Pattern record in Server Log

Internet: Five core protocols and many others terminology is used in Internet, which is given below.

- World Wide Web
- HTTP
- HTTP server information
- Client IP address or hostname

- URL

Web content mining is further divided into web page content mining and search mining.

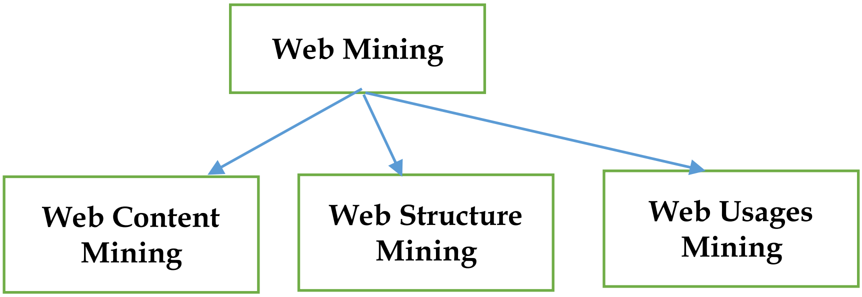

Fig. 2 Web Mining

Web Content Mining

Web content mining goes beyond basic IR technology. Web mining divided web content into agent based and database approach. Traditional mining operations involved:

- Keyword
- Term association
- Similarity search
- Classification
- Clustering
- Natural language processing

Web Structure Mining

Web structure mining can improve on the effectiveness of search engine and crawler.

- Page Rank
- Clever

Web Usages Mining

In order to discover usage patterns from the available data, it is necessary to perform three steps:

- Pre-processing
- Pattern discovery
- Pattern analysis

The most used techniques applied to Web usage data, as mentioned below:

- Statistical analysis
- Association rules
- Sequential patterns
- Clustering

Modes of Personalization

Personalization falls into four basic categories, ordered from the simplest to the most advanced:

- Memorization
- Customization
- Guidance or Recommender
- Task Performance Support

The Web personalization process can be divided into four distinct phases.

- Collection of Web data
- Implicit data
- Explicit data
- Preprocessing of Web data
- P address within a given time period.
- Analysis of Web
- Automatic user profiling

- Decision making/Final Recommendation Phase

Categories of Data used in Web Personalization

The Web personalization process relies on one or more of the following data sources.

- Content Data
- Structure Data
- Usage Data.
- User Profile

Challenges in WWW Personalization

WWW personalization faces several tough challenges that distinguish it from the main stream of data mining:

- Scalability
- Accuracy
- Evolving User.
- Data Collection and Preprocessing
- Integrating Multiple Sources of Data
- Conceptual Modeling for Web usage Mining
- Privacy Concerns

CONCLUSION

The various concept of the web mining techniques have been discussed in depth. The basic steps followed in web mining have been clarified in detail and also discussed personalization and its challenges.

REFERENCE

1. *Thomas W. Miller, Data and Text Ming*
2. *Margaret H.Dunham,S.Sridhar: Data mining*
3. *Bamshad Mobasher, Namit Jain, Eui-Hong Han, Jaideep Srivastava (1996). Web mining: pattern discovery from World Wide Web Transactions.*

4. *Robert Cooley, Bamshad Mobasher, Jaideep Srivastava (1997). Grouping Web page references into transactions for mining World Wide Web browsing patterns.*

5. *Robert Cooley, Bamshad Mobasher, Jaideep Srivastava (1997). Web Mining: information and pattern discovery on the World Wide Web.*

6. *Masseglia, P. Poncelet, M. Teisseire (1999). Using data mining techniques on Web access logs to dynamically improve Hypertext structure.*

7. *Robert Cooley, Bamshad Mobasher, Jaideep Srivastava (1999). Data preparation for mining World Wide Web browsing patterns.*

8. *Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang- Ning Tan (2000). WebUsage Mining: Discovery and applications of usage patterns from Web data.*

9. *Sudhakar Ranjan, Komal Kumar Bhatia "Query interface integrator for domain Specific Hidden web" International Journal of Computer Engineering and Applications (IJCEA) ISSN 2321-3469, Vol. IV, Issue I/III, Oct.-Dec. 2013.*

10. *Sudhakar Ranjan, Komal Kumar Bhatia "Indexing for Vertical Search Engine: Cost Sensitive"International Journal of Emerging Technology & Advanced Engineering (ISSN 2250-2459, ISO 9001:2008 Certified Journal), Volume 3, Issue 10, October 2013.*

11. *.Sudhakar Ranjan, Komal Kumar Bhatia"Web Log for Domain Specific Hidden Web", International Journal of Computer Engineering and Applications (IJCEA) ISSN 2321-3469, Vol VII, Issue I, July 2014.*

12. *Sudhakar Ranjan, Komal Kumar Bhatia "Indexing for Domain Specific Hidden Web" International Journal of Computer Engineering and Applications (IJCEA) 2321-3469, Vol VII, Issue I, July 2014.*

13. *Sudhakar Ranjan, Komal Kumar Bhatia "Transaction in Hidden Web" International Journal of Computer Engineering and Applications (IJCEA) 2321-3469, Vol VIII, Issue II, Nov 2014.*

14. *.Sudhakar Ranjan, Komal Bhatia: Design a Least Cost Vertical Search Engine based on DSHWC" selected for publication in IJIRR ,2017*

15. *Sudhakar Ranjan, Sarim Moin, Parikshit Vasisht, Abdus Samad Moin Uddin, "10 Role of Cyber-Security in Smart Energy Management Systems", Smart Energy Management Systems and Renewable Energy Resources, AIP publishing, Sepetember 2021 https://doi.org/10.1063/9780735422827_010*