JYOTI NIVAS COLLEGE

POST GRADUATE CENTER

DEPARTMENT OF MCA

III YEAR

TECH-ON-TOP

E-JOURNAL

ON

Issue 3, October 2020

SOCIAL NETWORK ANALYSIS IN WEB MINING

NIKHILA W (18MCA24)

MAMATHA N (18MCA09)

INTRODUCTION:

A social network is defined as a structure between individuals or organizations and their personal connections. It allows users to share ideas, activities, events, and interests within their individual networks.

Social networks are typically rich in text, because of a wide variety of methods by which users can contribute text content to the network.

Social network analysis maps and measures relationships and flows between people, groups, organizations, computers, URLs, and other connected information and knowledge entities.

Few of the social media platforms are Facebook, Twitter, LinkedIn, Instagram, Blogging, Wikis, emails and chat.

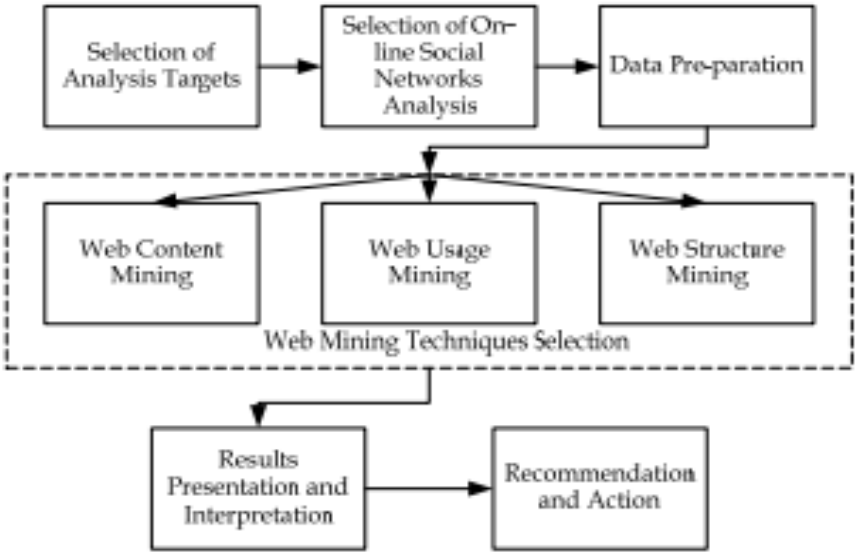
WEB MINING: -

Web mining is an application of data mining, the technique of discovering and extracting useful information from large datasets. Web mining techniques can be divided into three different types:

1. Web content mining analyses the content on the website, such as text, graphs etc. Most web content mining has focused on text data processing.
2. Web structure mining is a technique that analyses and explain the links and structure of websites.
3. Web usage mining can be used to analyse how websites have been used, such as determining the navigation behaviour of users.

THE PROCESS OF WEB MINING FOR SOCIAL NETWORKS ANALYSIS: -

The below figure presents a general process using web mining for social networks analysis. Its steps include selection of analysis targets, selection of on-line social networks analysis, data preparation, web mining techniques selection, results presentation and interpretation, recommendation and action.



- The first step is the selection of the analysis targets, such as websites, email, telephone communications etc.
- After this step, we select what of kind social networks analysis we will proceed with.
- The next step is data preparation. In this stage related data will be collected for analysis, then cleaned and formed as the final format to store in dataset.
- The next step is selecting the web mining techniques (web content mining, web usage mining and web structure mining to be used and then proceeding with them. More than one technique may be selected and sometimes a combination of techniques is necessary.
- The results of the analysis after web mining are then presented and interpreted either by manually or automatically. Visualization techniques are used to assist the presentation of the results of the analysis.
- The last step of the general process to use web mining for social networks analysis is recommendation and action.

REFERENCES

- https://ieeexplore.ieee.org/abstract/document/4598506
- https://www.slideshare.net/akash_mishra/data-mining-in-social-network

WEB MINING WITH SENSOR NETWORKS

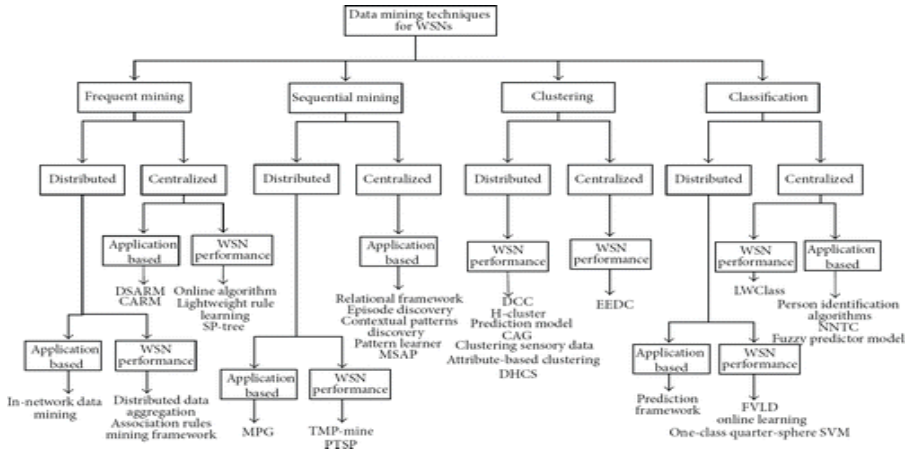
KUSHMETHA K. A(18MCA08)

BRUNDA S(18MCA05)

Web Mining is the process of Data Mining whose main purpose is discovering useful information from the World Wide Web and its usage patterns. It is used to understand customer behaviour and to evaluate the effectiveness of a particular website. It is especially used in Google,Yahoo and vertical searching like FatLens. Some special tools for Web Mining are Scrapy,PageRank and Apache logs. It is very useful for particular websites and e-services. The information gathered through Web Mining is evaluated using traditional Data Mining parameters.

Mining Techniques:

The highest-level of classification is based on general data mining classes.Some are frequent pattern matching and sequential pattern matching.The second-level is based on each approach's ability to process the data in centralized or distributed manner.This helps to improve the lifespan of sensor networks. The third level is selected according to the attribute towards solving specific problems. There are 2 separate aspects of issues: performance issue and application issue.



Challenges:

- Resource Constraint:- In terms of power,memory,communication bandwidth and computational power.
- Fast and Huge data arrival:- Data arrives faster than we are able to mine and may cause many classical data processing techniques to perform poorly.
- Online Mining:- Few techniques that analyze data in an offline manner do not meet the requirement of handling distributed stream data.

References:

https://journals.sagepub.com/doi/10.1155/2013/406316

https://www.geeksforgeeks.org/web-mining/

WEB STRUCTURE MINING

SHIVAVARSHNI.K (18MCA18)

SUMALATHA.M (18MCA20)

Introduction

The World Wide Web contains a large amount of information. Everyone can store and retrieve the information from web. Extracting the important information from web is called web mining. Web mining is one if the mining technologies which applies data mining techniques in large amount of web data to improve the web services. Web mining has three categories Web content mining, web usage mining, web structure mining.

Web Structure Mining

Web structure mining is the process of discovering information from the web. Web structure mining aims to generate structural summary about web sites and web pages. The focus of structure is therefore no link information, which is important of web data.

Types of Web Structure Mining

* Web graph mining

  Web is considered as graph which helps web algorithms for crawling, searching and community discovery, and the sociological phenomena which characterize its evolution.
* Web information extraction

  Web information extraction focuses on extracting structures with various accuracy and granularity out of Web pages. Web content structure is a kind of structure embedded in a single Web page and is also called intrapage structure.
* Deep Web mining

  It is hidden from users. The hidden part of web is called deep web or hidden web.

Techniques:

- Link based classification
- Link based cluster analysis
- Link type
- Link strength
- Link cardinality

Algorithm used:

- Page rank algorithm
- HITS algorithm
- Weighted page rank algorithm
- Distance rank algorithm
- Weighted page content rank algorithm
- Eigen rumour algorithm
- Time rank algorithm
- Query dependent ranking algorithm

Conclusion

Web structure mining is used to extract information from web pages. Search engine helps to obtain required information in an efficient and systematic manner. The search becomes user friendly. PageRank and Weighted PageRank algorithm analyzes only the link structure, whereas HITS algorithm gives some preference to web page content.

References

https://core.ac.uk/download/pdf/25814737.pdf

https://www.geeksforgeeks.org/web-mining/

http://www.cyberartsweb.org/cpace/ht/lanman/wsm1.htm

https://www.slideshare.net/AmirFahmideh/web-mining-structure-mining

WEBSITE EVALUATION USING OPINION MINING

AISWARYA S PILLAI (18MCA01)

RENI MATHEW (18MCA14)

INTRODUCTION:

We used opinion mining methods to evaluate various websites present on the internet. Also analyse the approaches, tools, and dataset used by Scholars with their accuracy. Opinion mining is hardly used in websites evaluation. Now a day's the websites we regularly uses are spamming with advertisements and unusable contents. By evaluating a website using the user feedback on the website collected on our website. That collected feedback data is processed by using a data mining software.

In Website Evaluation System it will rates the website based on opinions of the users. Website will be evaluated based on the factors such as genuineness of the website, timely delivery of the product after online transaction and support provided by the website. The user will comment about the website based on the comment system that is rating the website. This system takes an opinion of various users based on the opinion. The System will decide whether the website is genuine or not. The system uses opinion mining methodology for mining, in ordered to achieve the desired functionalities. We will use a database for sentimental based keywords along with positivity or negativity weight in the database and also based on the sentimental keywords mined in the user comment is ranked. The system contains keywords which is related to fraud, genuineness, timely delivery of the product and service meters in the database. Based on these factors system will rate the website.

THE WORKING OF SYSTEM AS FOLLOWS:

- The users log in to the system, so he can view various websites posted by the admin and also can comment about the website.
- User is able to see the comment of the other users.
- The System will help to rate the website based on the comment of various users.
- The role of the admin is to add various website to the system and to add keywords in database.
- Through this system will match the comment with the keywords in database, it will rate the website based on the sentimental analysis.
- Users can easily find out which website will deliver the product on time. And also helps to find out website that, which will provide the good support.
- This application will help to find out whether the website is genuine or not, that is very useful for those users who do the online transactions.

ADVANTAGES:

- User can easily share their views about the website.
- People can easily find whether the website is genuine or not
- More useful for those who do the online transactions.
- Helps the user to find out the website, which provides good service and delivers the product on time.
- System ranks the website based on the weight age of the keywords in database so the result is appropriate.

DISADVANTAGES:

- The System will match the opinion with those keywords which are in the database rest of the words are ignored by the system itself.

REFERENCE:

https://nevonprojects.com/website-evaluation-using-opinion-mining/

https://www.sciencepubco.com/index.php/ijet/article/view/10257

WEB INFORMATION RETRIEVAL

MADHU SHREE S M

(182MCA27)

"Web information retrieval" is the process of searching useful information within world wide web document collection.

Characteristics of web information retrieval

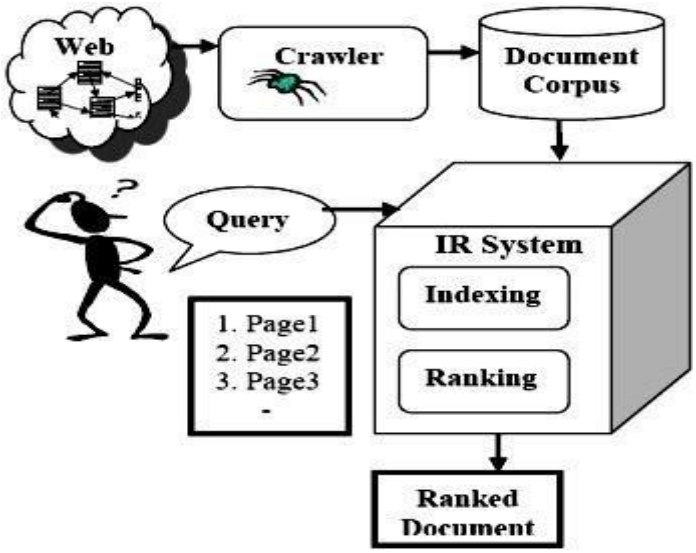
- **Dynamic** : The web is dynamic . The information on the web will be constantly changing and being updated.
- **Heterogeneity** : The data is heterogeneous , which exist in multiple languages , formats etc..
- **Huge size :** The web is vey huge in size.
- **Self-organized :** The data will be collected and organized by trained specialist.

Components of Web IR

- **Crawler :** The crawling software will create virtual spiders , which will scour the new information and web pages and will store in central repository.
  Ex : Googlebot
- **Page repository :** The spiders will return new web pages , which are temporarily stored in page repository.
- **Indexing Module :** This module will take uncompressed page and extract only vital information and creates compressed page that is stored in various indexes.
- **Ranking module :** This module will take the set of relevant pages from query module and will rank them based on some criteria.

Types of Web IR Users

- **Casual User :** searching the web for something that is loosely defined.
- **Researcher :** searching the information for research level in the web.
- **Professional :** looking for the business intelligence by searching the web.



Web IR Tasks

- **Filtering :** Selecting the document using fixed query in dynamic collection.
- **Ad-Hoc Retrieval :** This is the standard retrieval task in web.
- **Clustering :** Grouping the documents in pre-defined structure.
- **Topic distillation :** Finding the short list of points.
- **Homepage finding :** Finding the URL of named entity.

Web IR Tools

- Search Tools
- Search Services

Conclusion

Web information retrieval is one of the rapidly growing technology in today's world.AS the we keeps growing in size,the problem of searching the web becomes more complex. Web IR can be defined as the application of theories and methodology from IR to WWW..

References

- https://link.springer.com/referenceworkentry/10.1007%2F978-1-4614-8265-9_928
- https://www.springer.com/gp/book/9783642393136
- https://www.researchgate.net/publication/289907929_web_information_retrieval

WEB MINING INTEGRATE WITH ARTIFICIAL NEURAL NETWORKS.

SUDESHNA BOSE(18MCA19)

SWATHI(18MCA21)

The Application of data mining techniques to the World Wide Web referred as web mining. When we see web mining in terms of data mining it have three interest of operations ie clustering ,associations and sequential analysis.

ANN is an information processing technology. applications of ANN are pattern recognition and classification .it plays a major role in web mining. Nonlinearity, Adaptivity, Input-output mapping, contextual information these are some features of artificial neural networks.

Web mining techniques can be categorized as web content mining, web structure mining and web usage mining.

Web-content mining: Its all about search and retrieval of information on the web. Web content may be unstructured (plain text), semi-structure (HTML documents), or structured (extracted from databases into dynamic web pages).

Web-structure mining: The goal of web structure mining is to categorized the web pages and generate the information such as the similarity and relationship between them, taking the advantage of their hyperlink topology.

Web-usage mining: It discovers and analyzes user access patterns. Web usage mining is the process of identifying browsing patterns by analysing the user's navigational behaviour.

Some of the Web Mining Applications:

- E-Commerce
- Search Engines and Web search
- Website Design
- Recommendation engines
- Web communities and web market places.

Some Issues on Web Mining:

- Web data sets can be very large
- Cannot mine on a single server
- How to organize hardware and software to mine multi-tera byte data sets?

Web mining enables us to screen specific data through content mining, to discover the structural summary of web sites through structure mining and to predict the behaviour and interaction of the surfers with the web through usage mining.

References:

- https://www.sciencedirect.com/science/article/abs/pii/S0957417409008288
- http://www.ijesi.org/papers/RTSCA-2K17-2017/B1216.pdf
- https://www.researchgate.net/publication/222690840_Integrating_web_mining_and_neural_network_for_personalized_e-commerce_automatic_service.

RECOMMENDER SYSTEMS AND COLLABORATIVE FILTERING

AFRIN SULTANA

(182MCA43)

Web mining is the process of Data Mining techniques to automatically discover and extract information from the web document and services. The main purpose of web mining is discovering useful information from www and its usage pattern.

Recommender systems and collaborative filtering

Collaborative filtering is the process of filtering or evaluating items using the opinions of other people. It is one of the core technologies that will power the adaptive web. In order to filter information based on such complex dimensions, we need to include people in the loop, who analyze the information and condense there opinion into data that can be easily processed by software.

Algorithms used :

- Non-probabilistic algorithm
- probabilistic algorithm
- user-based nearest neighbor algorithm

Advantages

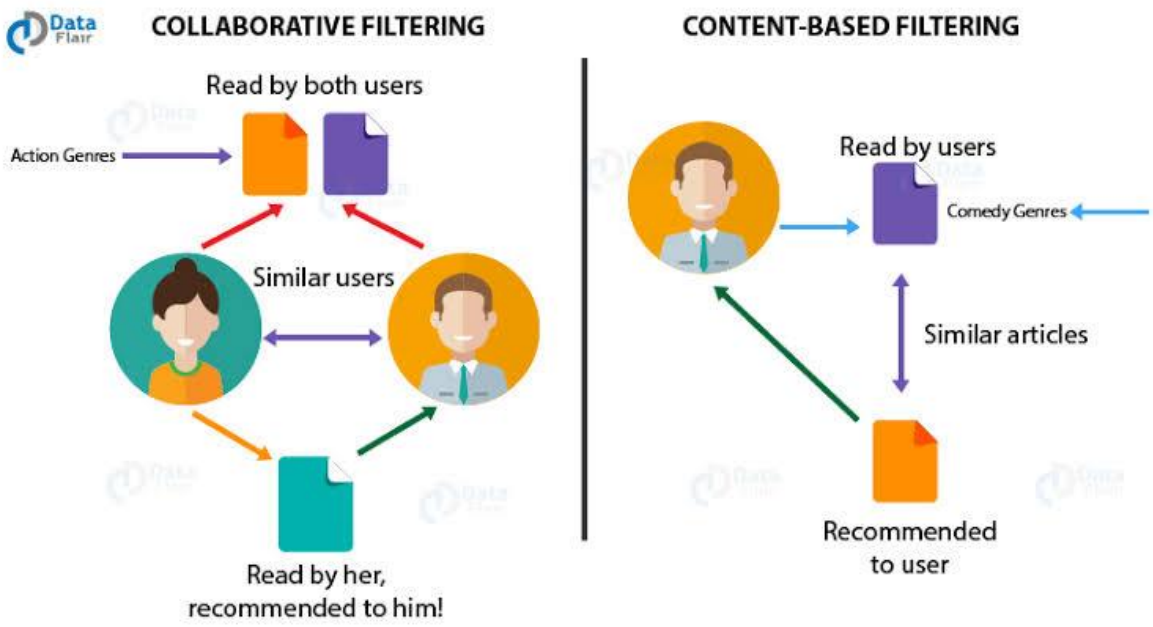
- No domain knowledge necessary
- Great starting point

Disadvantages

- Cannot handle fresh items.
- Hard to include side feature for item.

Types of collaborative filtering

- User-based, which measures the similarity between target user and other users.
- Item-based, which measures the similarity between the items that target users rate or interact with and other items.



Implementations of recommendation system

- Collects and organize information on users and products.
- Compare user A to all other users.
- Rank and Recommend.
- Evaluate and test.
- Wrapping up.

References

1. Avery, C., Resnick, P., Zeckhauser, R.: The Market for Evaluations. American Economic Review 89(3), 564–584 (1999)
2. Balabanovíc, M., Shoham, Y.: Fab: Content-Based, Collaborative Recommendation. Communications of the ACM 40(3), 66–72 (1997)

WEB CONTENT MINING

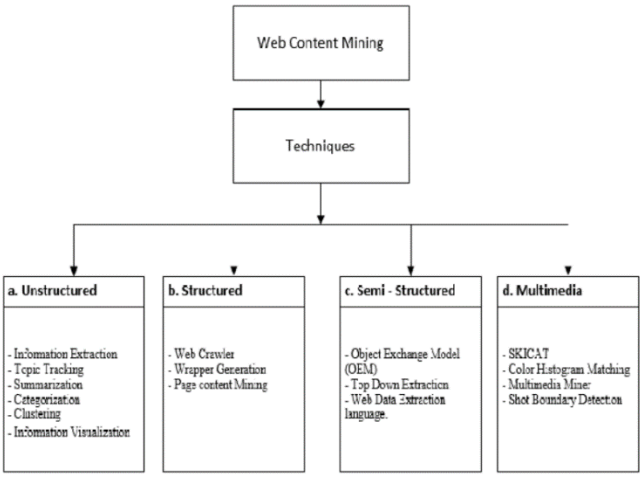
ANKITHA.K.V(182MCA30)
BRUNDA.V(182MCA35)

Web content mining

The extraction of certain information from the unstructured raw data text of unknown structures is referred to as **Web content mining**. A set of information extraction tools is brought forward in order to identify and collect **content** items, such as Text Extraction and Wrapper Induction. Web content mining targets the knowledge discovery, in which the main objects are the traditional collections of multimedia documents such as images, video, and audio, which are embedded in or linked to the web pages.

Web Content mining has approaches to mine data: unstructured mining, structured mining, semi-structured mining and multimedia mining.

Web content mining Techniques:



Web content mining Tools:

Web content mining tools are software that helps to download the essential information for users as it collects appropriate and perfectly fitting information. Some of the tools are:

A. Web Info Extractor (WIE): This is a tool for data mining, extracting Web content, and web content Analysis and it can extract structured or unstructured data from Web page, reform into local file or save to database, place into Web server.

B. Mozenda: This is a tool to enable users to extract and manage Web data. The Users can setup agents that normally extract, store, and also publish data to multiple destinations.

C. Screen-Scrapper: This is a tool for extracting/mining information from web sites. It is used for searching a database, which interfaced with software to attain content mining needs. The programming languages such as Java, .NET, PHP, Visual Basic

D. Web Content Extractor (WCE): WCE is a powerful and easy to use data extraction tool for Web scraping, and data extraction from the Internet.

Web Content Mining Problems/Challenges:

- Data/Information Extraction: Extraction of structured data from Web pages, such as products search results is a difficult task.
- Opinion extraction from online sources: There are many online opinion sources, like customer reviews of products, forums, blogs and chat rooms.
- Knowledge synthesis: Concept hierarchies or ontology are useful in many applications. However, generating them manually is very time consuming. A few existing methods that explores the information redundancy of the Web will be presented.
- Segmenting Web pages and detecting noise: In many Web applications, one only wants the main content of the Web page without advertisements, navigation links, copyright notices.

References:

[1] P. Maes. Agents that reduce work and information overload. Communications of the ACM, 37(7):30–40, 1994.

[2] M onik a Yad a v , Pra d eep Mi tt a l , " Web Mining: An Introduction ", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013 ISSN: 2277 128X.

[3] Johnson, Faustina, and Santosh Kumar Gupta. "Web Content Mining Techniques: A Survey."International Journal of Computer Applications (0975–888) Volume (2012) [4] Sharma, Arvind Kumar, and P. C. Gupta. "Study & Analysis of Web Content Mining Tools to Improve Techniques.

WEB USAGE MINING

TKPS PRAGNA VALLIKA (182MCA29)
RAKSHITHA V (182MCA36)

The web usage mining deals with identifying the patterns of user, client and server interaction with one another. It focuses on techniques that have the potential to predict user behaviour while the user interacts with the Web followed by various applications as well. Web usage mining is the part of web mining that helps to retrieve information of user interaction from the user or client server.

The sources of data from where this extracts information is from Server access logs, agent logs, refer logs, cookies, User profiles, User ratings, Click streams, Database transactions on websites, Page content and site structure so on. The web mining consists of four essential phases. They are the data collection, data pre-processing, pattern discovery and pattern analysis.

Data collection:

In data collection there are three main sources for data in usage mining: server-side data, client-side data and intermediary data. Server data are data that are collected from web servers; it includes log files, cookies and explicit user input. Cookies are strings that are sent from the web server to the client's browser. The browser saves the cookie in a text file and resends it to the server each time they visit the site. The last source of data is intermediate data, which can be collected from proxy servers or packet sniffers.

Pre-processing:

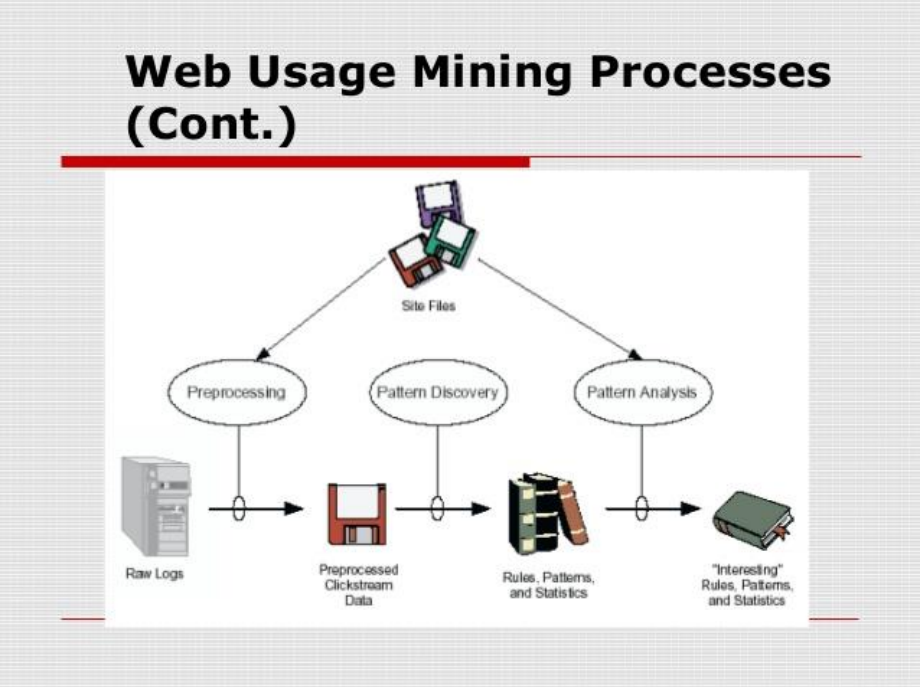
Data preparing includes data cleansing where any unnecessary items should be removed; for example, videos, audio files, graphics and the format information, meaning files that have particular suffixes (JPEF, GIF, etc.) which are usually downloaded without user consent. User identification is to identify who accesses the website and which pages are accessed, in session identification which encodes the navigational behaviour of the users, is very important in usage mining and path completion where the missing pages path should be added using different techniques.

Pattern discovery:

In pattern discovery the pre-treated information is analysed to extract valuable patterns. Statistical methods and machine learning are used to mine patterns. The better-known approaches used are path analysis, association rules, clustering, classification, sequential patterns and order model discovery.

Pattern analysis:

In pattern analysis After usage patterns are discovered, techniques and tools are needed to make these patterns understandable for analysts and to maximize the benefits from these patterns. Techniques include database querying, graphics and visualization, statistics and usability analysis.



Applications and tools used:

WUM – SiteHelper – WebPersonalizer WebSIFT – SETA –Tellim – Oracle9iAS Netmind – Litezia – Web Watcher Krishnapuram – Analog – Accure G2 – Accure Insight 5 – Pilot HitList – SEWeP are used for Personalization. SurfAid – Tuzilin – Burchner – SAS Intellivisor ECOMMINER – InterShop – Logisma Business Wbstore is used for Business intelligence and Etzioni – Perkowitz – iJADEMiner for Site design support.

Conclusion:

web usage mining is of great importance in many areas, especially in e-commerce applications and website designing.

References:

https://www.researchgate.net/publication/282837445_Application_and_Significance_of_Web_Usage_Mining_in_the_21st_Century_A_Literature_Review

WEB MINING TOOLS

DIVYA P(18MCA23)

SANGEETHA. G(18MCA16)

Introduction

A web mining tool is computer software that uses data mining techniques to identify or discover patterns from large data sets. Data is money in today's world, but the information is huge, diverse and redundant. Having the tools for mining is going to be a gateway to help you get the right information.

Different Web mining Tools are:

1. **R**
   R Language is used for statistical computing and Graphics . It is accessible from scripting languages like python, Ruby, Perl etc. Unix platforms supported.
   Area of web mining: Web usage mining
2. **Octoparse**
   It helps you create highly accurate extraction rules. Crawlers which run in Octoparse are determined by the configured rule. Windows platforms supported.
   Area of web mining: Web content mining.
3. **Oracle Data mining**
   Oracle Data mining is implemented in Oracle Database kernel and mining models are best database objects which uses built-in features of it. Windows platforms supported.
   Area of web mining: Web usage mining
4. **Tableau**
   Tableau gives fast response by transforming data into visually appealing, interactive-visualizations called dashboards and time taken is seconds or minutes rather than months or years. Windows and Mac platforms supported.
   Area of web mining: Web usage mining
5. **Scrapy**
   It's an opensource framework for collecting data from websites, written in python and rules must be written to extract web data. Windows, Linux, BSD and Mac platforms supported.
   Area of web mining: Web content mining
6. **Page Rank Algorithm**
   It is a link analysis algorithm and it assigns a numerical weighting to each element of hyperlinked set of documents to measure its relative importance within the set.
   Area of web mining: Web structure mining

References

1. https://prowebscraper.com/blog/web-mining-tools/
2. https://www.octoparse.com/blog/7-web-mining-tools-around-the-web
3. https://en.wikipedia.org/wiki/Web_mining

IMAGE MINING

CELESTINE JEENA (18MCA06)
S. LEISHIPEM SOPHIA (18MCA15)

Introduction: Image mining is a well-structured technique based on data mining, artificial intelligence, machine learning, image retrieval, image processing, computer vision and database etc. Image mining aims to extract relationships and patterns which are not explicitly stored in database from raw data images. To use them in high-level modelling they must be processed first. An image mining technique is considered as a good technique if it supports fully user interaction during retrieving the patterns and knowledge from the collection of huge image database. In the last decade, data mining as a research field has expanded and progress in data processing is getting both more accurate and convenient.

Image mining techniques

The techniques frequently used are classified on five levels of information and the associated image or data mining operations. These levels are-

- knowledge extraction level
- patterns and inter-image relations level
- semantic concept level
- region, objects, or visual patterns level
- pixel level

The techniques used are as follows-

1. *Object recognition*- One of the key areas which operates data on patterns and inter-image relations level. It finds the object relevant to the real world, from the image by processing the provided object models. It is also known as supervised labelling method. The system has four parts, they are:
    a. feature detector
    b. model database
    c. hypothesiser
    d. hypothesis verifier.
2. *Image retrieval*- refers to the process of retrieving a particular image from a large database using data mining. Retrieval of images in image mining (Tahoun et al., 2005) is done based on some requirement specification. There are three levels of requirement specifications and the complexity also increase with the levels.
    - level 1 retrieve the image based on some basic features of images such as texture, colour, shape or image elements' spatial location
    - level 2 is based on image retrieval which derives the logical features such as individual objects or persons from images
    - level 3 is based on image retrieval by abstract attributes which involves a high level reasoning in order to obtain the meaning of the objects or scenes illustrated.
3. *Image indexing*- Apart from focusing on the information requirements at various levels, it is also important to provide support for the retrieval of image data with a fast and efficient indexing scheme. Image indexing handles data and images in region, objects and visual patterns level. Reducing the dimensions can be accomplished using two well-known methods:
    a. the singular value decomposition (SVD) update algorithm
    b. clustering.
4. *Image classification and clustering*- Supervised and unsupervised classification of images into groups respectively is basically performed on the basis of classification and

clustering. The main motive for performing classification or clustering on images is to retrieve knowledge that the users demand from the image group stored in the image database associated with the image. In supervised classification, the pool or set of labeled images which are used to label newly unlabeled images that are encountered. In image clustering, the main motive is to combine a number of unlabeled images into meaningful groups called clusters. Based on image content, these clusters are formed without any prior knowledge.

5. *Association rule mining-* is a typical approach used in data mining domain for uncovering interesting trends, patterns and rules in large datasets. An association rule is an implication of the form X® Y, where X, YÌ I and XLY=f. I is the set of objects, also referred as items. D is a set of data cases. X is called the antecedent and Y is called the consequent of the rule. A set of items, the antecedent plus the consequent, is call an itemset.

Image mining frameworks- Function-driven frameworks and Information-driven frameworks

Image mining real-world application

- Monsoon and Rainfall Prediction: The prediction of rainfall is one of the major studies in field of image mining. In India, where agriculture is major occupation which is dependent on rainfall, the time and amount of rainfall holds high importance and thus mining useful data relevant to this is on high demand.
- Satellite Image Mining: The satellite images contain information for weather forecasting and early prediction of different natural calamities like typhoon, hurricanes etc. Other parameters like humidity, linear cloud, typhoon can be extracted from satellite images to get a useful and efficient knowledge.
- Textile Image Retrieval Using Color as parameter: All textile industries aim to produce large scale of textile depends mainly on designs and quality of the dresses produced. Every day, numerous textile images are being generated such as images of shirts, jeans, t-shirts and sarees. Images play an important role as a picture is worth thousand words in the field of textile design and marketing. A retrieving of images needs special concepts such as image annotation, context, and image content and image values.

Conclusion- Image mining is a progressive field that retrieves images that best matches from the images in the image dataset, on the basis of query image. It is a new and promising area for knowledge extraction from images. We have discussed techniques that are frequently used in the early works in image mining and the real-world applications.

References

[1] A Study on Image Mining Methods and Techniques Ankita Tripathi1, Hitesh Jangir2 International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 4, Issue 4, April 2016

[2] Image Mining: Issues, Frameworks and Techniques Ji Zhang, Wynne Hsu, Mong Li Lee Department of Computer Science, School of Computing National University of Singapore

[3] Image mining framework and techniques: a review, Article in International Journal of Image Mining · April 2015

[4] SURVEY OF IMAGE MINING TECHNIQUES AND APPLICATIONS R. Vijayalatha Research Scholar, Manonmaniam Sundaranar University, Tirunelveli (India) February 2017

[5] Image Mining: Review and New Challenges (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 6, No. 7, 2015

WEB MINING IN HEALTH CARE

PREETHI C (182MCA41)

GAYATHRI K S (182MCA40)

Web Mining is the process of data mining techniques which will automatically detect and extract the information from the web documents and services. Web Mining also plays a prominent role in predicting the user behavior.

Web Mining is known for its richness in having applications in variety of domains like corporate, e-learning health care etc. There are enormous number of health care websites and online services which provide information about symptoms and cause of the disease, its cure and side effects. We can collect clinical data from various websites about a particular disease and predict the pattern with the help of web mining techniques. **Health care web mining**, a field developed to improve the methods that makes use of the knowledge that is acquired from the medical environment. The data obtained from the databases of the various medical institutes or any doctor's clinic, is used to understand patients behavior and diseases to assist the doctor. As a result to improve the health care systems and provides treatments with reduced errors or zero error treatments.

Health care web mining makes use of various techniques such as Neural Networks, K-nearest neighbor, decision trees and many more. These methods are utilized to acquire few kinds of knowledge such as classification, clustering. The data acquired can be used by an organization to identify a particular disease in a region, how many patients will register for a given disease. Web mining helps to identify hidden patterns in medical institutes' database. These patterns are extracted to build mining models and in turn used to predict the diagnosis and assist the doctors in taking the decisions with high accuracy. Therefore hospitals are able to allocate resources more effectively. As a result of this, a hospital will be able to take necessary precautions before the patient decides to quit their treatment, or to efficiently assign resources with an accurate estimate of how many male or female will register for a particular disease by using the Prediction techniques.

REFERNCES

[1]. An Empirical Study of the Applications of Web Mining Techniques in Health Care Dr. Varun Kumar Department of Computer Science & Engineering ITM University, Gurgaon, India MD. Ezaz Ahmed Department of Computer Science & Engineering ITM University, Gurgaon, India

[2]. https://www.geeksforgeeks.org/web-mining/

Multimedia Data Mining

NAVYA P
(18MCA12)

Abstract:

In digital media technologies has made transmitting and storing large amounts of Multimedia data such as text, images, music, video and their combinations of several types are becoming increasingly available and are almost unstructured or semi structured data by nature, which makes it difficult for human beings to extract the information without powerful tools. more feasible and affordable than ever before. Multimedia mining deals with the extraction of implicit knowledge, multimedia data relationships. This drives the need to develop data mining techniques that can work on all kinds of data such as documents, images, and signals.

INTRODUCTION:

- Multimedia data mining is used for extracting interesting information for multimedia data sets.
- Multimedia mining is a subfield of data mining which is used to find interesting information implicit knowledge from multimedia databases.
- **Text Data:**
  It is used to find meaningful information from the unstructured texts that are from various sources.
- **Image Data :**
  Image data system can discover meaningful information or image patterns from a huge collection of images.
- **Video Data:**
  Video data is unsubstantiated to find the interesting patterns from large amount of video data.
- **Audio data:**
  Audio data plays an important role in multimedia applications , is a technique by which the content of an audio signal can be automatically searched, analyzed and rotten with wavelet transformation.

Feature Extraction from Color Images:

Image categorization classifies images into semantic databases that are manually re categorized.

Three types of feature vectors for image description:

1) Pixel level features,
2) Region level features and
3) Tile level features.

APPLICATIONS OF MULTIMEDIA MINING :

There are different kinds of application of multimedia data mining some of which are as following:

- Digital library
- Traffic video sequences
- Medical analysis
- Customer perception
- Media making and broadcasting
- Surveillance system

Converting unstructured data to structured data :

- Data resides in fixed field within a record or file is called structured data and these data are stored in sequential form. Structured data has been easily entered, stored, queried and analyzed.
- Unstructured data is bit stream, for example pixel representation for an image, audio,video and character representation for text.
- These sorts of files may have an internal structure, they are still considered unstructured because the data they contain does not fit neatly in a database.

ARCHITURES FOR MULTIMEDIA DATA MINING:

Multimedia mining architeture has several components are:

- Input
- Multimedia content
- Finding the similar patterns
- Evaluation of results
-

CONCLUSION:

The multimedia mining, non-structured heterogeneous information: audio, video, image, speech, text, graphics, icons, web logs, etc. The multimedia mining, knowledge extraction in multimedia data mining. Multimedia data mining techniques are active and growing area.

REFERENCES:

1. Pravin M. Kamde1, Dr. Siddu. P. Algur "A SURVEY ON WEB MULTIMEDIA MINING" The International Journal of Multimedia & Its Applications (IJMA) Vol.3, No.3, August 2011
2. Manjunath T.N1, Ravindra S Hegadi2, Ravikumar G K "A Survey on Multimedia Data Mining and Its Relevance Today" IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.11, November 2010

TEXT MINING

SANIYA KULSUM (182MCA50)

KAVYA S (182MCA51)

Text mining, also referred to as text data mining, similar to text analytics, is the process of deriving high-quality information from text. Text mining is the process of examining large collections of documents to discover new information or help answer specific research questions.

Text mining identifies facts, relationships and assertions that would otherwise remain buried in the mass of textual big data. Once extracted, this information is converted into a structured form that can be further analyzed, or presented directly using clustered HTML tables, mind maps, charts, etc. Text mining employs a variety of methodologies to process the text, one of the most important of these being Natural Language Processing (NLP).

The structured data created by text mining can be integrated into databases, data warehouses or business intelligence dashboards and used for descriptive, prescriptive or predictive analytics.
Text mining methods and software is also being researched and developed by major firms, including IBM and Microsoft, to further automate the mining and analysis processes, and by different firms working in the area of search and indexing in general as a way to improve their results. Within public sector much effort has been concentrated on creating software for tracking and monitoring terrorist activities. For study purposes, Weka software is one of the most popular options in the scientific world, acting as an excellent entry point for beginners.

The five fundamental steps involved in text mining are:

- Gathering unstructured data from multiple data sources like plain text, web pages, pdf files, emails, and blogs, to name a few.
- Detect and remove anomalies from data by conducting pre-processing and cleansing operations. Data cleansing allows you to extract and retain the valuable information hidden within the data and to help identify the roots of specific words.
- For this, you get a number of text mining tools and text mining applications.
- Convert all the relevant information extracted from unstructured data into structured formats.
- Analyze the patterns within the data via the Management Information System (MIS).
- Store all the valuable information into a secure database to drive trend analysis and enhance the decision-making process of the organization.

Technology premise of Text Mining:

1. Summarization: It is the process of creating a summary of any document which consists of a large amount of information while the theme or main idea of a document is maintained.

2. Information Extraction: It is the process of utilizing relations within the text format. It uses pattern matching format.

3. Categorization: Categorization is the supervised learning technique which places the document according to content. Document categorization is largely used in libraries.

4. Visualization: Visualization is computer graphics used to represent information and visualizing relationships. It is beneficial to depict a more clearer output.

5. Clustering: Clustering involves document's textual similarity based on the unsupervised technique used for data analysis to divide the text into a manual exclusive group.

6. Question Answering: It includes natural language queries with questions and answers and finding an appropriate solution from the list of patterns.

7. Sentiment Analysis: Sentiment analysis is also known as opinion mining which is configured based on user's emotion with various categories such as positive, negative, neutral and mixed. It is used to get people's view and attitude towards anything with services and products.

References:

1.https://en.m.wikipedia.org/wiki/Text_mining

2.https://www.upgrad.com/blog/what-is-text-mining-techniques-and-applications/

WEB INFORMATION CLASSIFICATION AND CLUSTERING

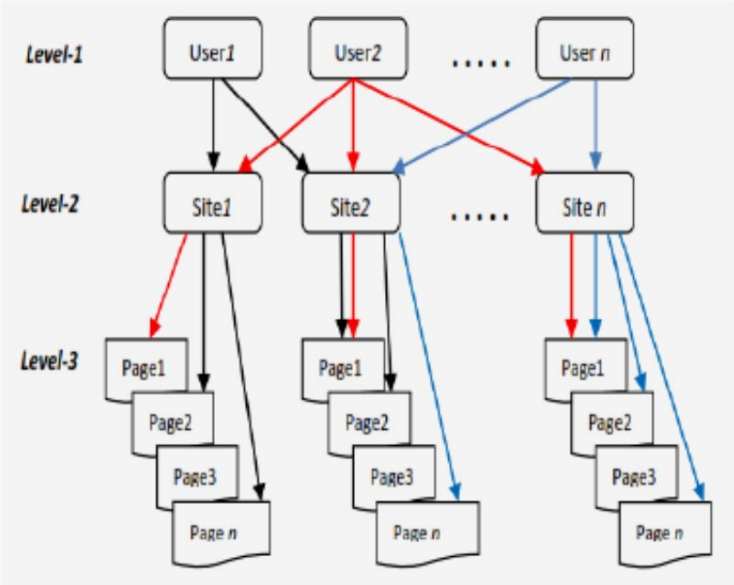
PAVITHRA D(18MCA13)
CHITHRA N(18MCA07)

Introduction:

In general search engine has the similar retrieved results to different type of users when submitting the similar type of query, no considering about other information needs and preferences.

Classification:

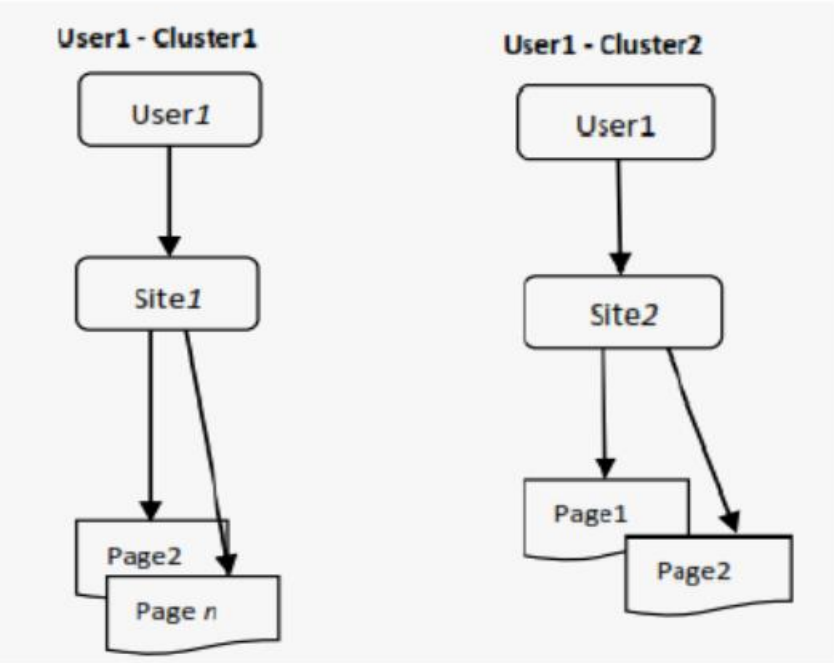
Classification applied to detect patterns using data mining methods. It helps data to extract data based on some rules. A multilevel association rule mining is applied on the proposed framework to build pattern on pre-processed data. Multi-level association rules can mine data log efficiently with the use of hierarchy under the support and trust framework. Overall, all top down strategy is employed, which counts accumulated to calculate the association item approach at all levels, from level 1.2 and 3concepts and working with hierarchy to more detailed conceptual level until you find related item sets.



Clustering:

Clustering group is similar and dissimilar to one another on the basics of other data, the same group in the group listing process of data. A cluster data that can be treated collectively as a group, and thus can be considered as a form of data compression. separate classification groups are effective tool, but in large set of models the characterize the sample in each group, in spite the need for proper collection and labelling.

The top-down manner in the framework of a hierarchical clustering method to group data into groups to implement a tree. This top-down strategy, starting with a cluster of top items, and each item satisfies the conditions to cancel some of its own to form a cluster, a cluster of tiny pieces and subdivides. web search can personalization on the back that they can monitor cluster samples stored in the ranking process.



References:

1. https://www.researchgate.net/publication/289527252_Web_Usage_Classification_and_Clustering_Approach_for_Web_Search_Personalization.
2. https://www.researchgate.net/publication/228374867_Clustering_techniques_utilized_in_web_usage_mining.

HYBRID CLUSTERING METHODS FOR WEB USAGE MINING

YOGESHWARI E (182MCA26)
MANISHA BURAGOHAIN (182MCA10)

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services.

There are three general classes of information that can be discovered by web mining:

» Web activity, from server logs and Web browser activity tracking.

» Web graph, from links between pages, people and other data.

» Web content, for the data found on Web pages and inside of documents.

Four Steps in Content Web Mining:

When extracting Web content information using web mining, there are four typical steps.

1. Collect – fetch the content from the Web

2. Parse – extract usable data from formatted data (HTML, PDF, etc)

3. Analyze – tokenize, rate, classify, cluster, filter, sort, etc.

4. Produce – turn the results of analysis into something useful patterns.

Clustering:

The most common techniques used for pattern discovery are clustering methods. Clustering is a way to divide a dataset into groups that differ from each other but whose elements are similar. There are two types of cluster to be discovered: page cluster, which aims to find pages of similar content, and usage cluster, which aims to define a group of users who have similar browsing patterns.

1)Web Usage Mining:

This study of the users' navigations extracted from the web server's log files or proprietary traces may help the webmaster to understand the user behavior and then to rethink the structure and design of his/her website or to detect users' problems and improve the navigability. The WUM analysis, allows the webmaster to optimize the response of the Web server (Web caching) and to make recommendations to the user.

2) Web Usage Data

The Web log file is the input data in the Web Usage Mining process. The Web site structure (hyperlinks graph) and the users' profiles may constitute supplementary data for such a process.

3) **Pre-processing**

The objective of the pre-processing step is to identify and structure user navigations. This step is based on two main processes: data cleaning and data transformation. Pre-processing of the Web log files Data Cleaning, Data Transformation, User/Session Identification

4) Document Clustering:

In this work, content mining is used approach for document clustering. Assume G = { g1, g2,..,gn }is the set of n website's pages.

1. Clear each document from stop words such as: about, all , am, almost, as, be, by, but, do and any other word which have not any key role in determining the content of document.

2. Identify document keywords by TF-IDF technique.

3. Assign each document keyword list as a document to a single cluster.

4. Merge primary clusters based on the Jaccard coefficient similarity measure.

5. The second step repeated until all documents being clustered into a pre defined number of clusters. DC = {DC1, DC2,…, DCn} is the result set. Each DCi represents a set of URLs with similar content.

CONCLUSION :

This model analysis the users behaviors and depend on the interests of similar patterns provides appropriate recommendations for active user. The model uses the benefits of both content based and collaborative based recommender systems. Proposed work can be extended by considering the effect of users feedback for increasing the quality of recommendation.

REFERENCES :

https://ieeexplore.ieee.org/document/1505425

ASSOCIATION RULES AND SEQUENTIAL PATTERNS

DINAMANI R(182MCA33)

MEGHANA R(182MCA32)

Association Rules

Association rules are an important class of regularities in data. Mining of association rules is a fundamental data mining task. It is perhaps the most important model invented and extensively studied by the database and data mining community. Its objective is to find all co-occurrence relationships, called associations, among data items. Since it was first introduced in 1993 by Agrawal et , it has attracted a great deal of attention. Many efficient algorithms, extensions and applications have been reported. The classic application of association rule mining is the market basket data analysis, which aims to discover how items purchased by customers in a supermarket (or a store) are associated.

The problem of mining association rules can be stated as follows:

Let I = {i1, i2, …, im} be a set of items. Let T = (t1, t2, …, tn) be a set of transaction (the database), where each transaction ti is a set of items such that ti $\subseteq$ I. An association rule is an implication of the form,

$X \rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$.

X (or Y) is a set of items, called an itemset.

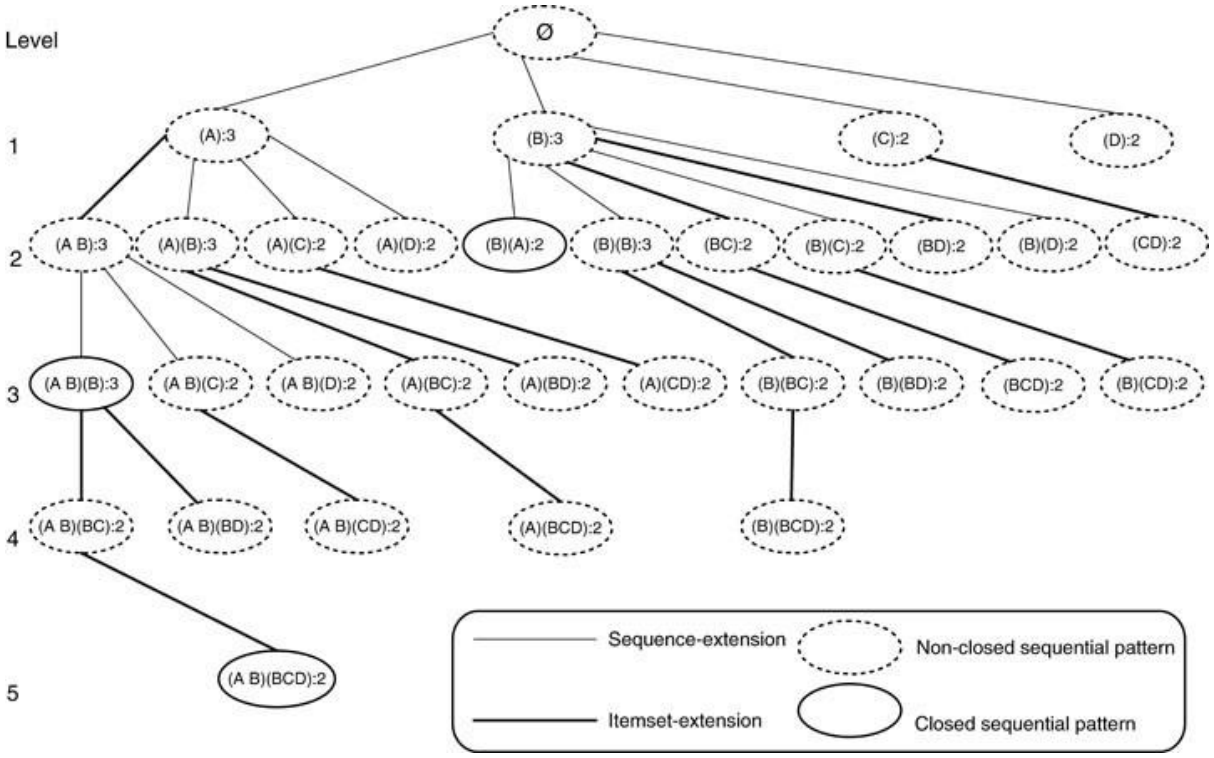
Sequential Pattern Mining

Association rule mining does not consider the order of transactions. However, in many applications such orderings are significant. For example, in market basket analysis, it is interesting to know whether people buy some items in sequence, e.g., buying bed first and then buying bed sheets some time later. In natural language processing or text mining, considering the ordering of words in a sentence is vital in finding language or linguistic patterns. For such applications, association rules are no longer appropriate. Sequential patterns are needed. Sequential patterns have been used extensively in Web usage mining for finding navigational patterns of users in the Web site. They have also been applied to finding linguistic patterns for opinion mining.

Problem Definition

Let I = {i1, i2, …, im} be a set of items. A sequence is an ordered list of itemsets. Recall an itemset X is a non-empty set of items X $\subseteq$ I. We denote a sequence s by ⟨a1a2…ar⟩, where ai is an itemset, which is also called an element of s. We denote an element (or an itemset) of a sequence by {x1, x2, …, xk}, where xj $\in$ I is an item. We assume without loss of generality that items in an element of a sequence are in lexicographic order. An item can occur only once

in an element of a sequence, but can occur multiple times in different elements. The size of a sequence is the number of elements (or itemsets) in the sequence. The length of a sequence is the number of items in the sequence. A sequence of length k is called k-sequence. If an item occurs multiple times in different elements of a sequence, each occurrence contributes to the value of k. A sequence s1 = ⟨a1a2…ar⟩ is a subsequence of another sequence s2 = ⟨b1b2…bv⟩, or s2 is a supersequence of s1, if there exist integers $1 \leq j1 < j2 < \ldots < jr-1 < jr \leq v$ such that $a1 \subseteq bj1$, $a2 \subseteq bj2$, …, $ar \subseteq bjr$. We also say that s2 contains s1.



References :

1)Zijian zheng,Real world performance of association rule algorithmsJanuary 2001
DOI: 10.1145/502512.502572 Source DBLP

2) Hussam Sherman, An Efficient Algorithm for Mining Association Rules for Large Itemsets in Large Databases April 2014

3) Advanced Topics on Association Rules and Mining Sequence Data Lecturer: JERZY STEFANOWSKI Institute of Computing Sciences Poznan University of Technology Poznan, Poland Lectures 11 SE Master Course 2010

OPINION MINING AND SENTIMENT ANALYSIS

DEEKSHITHA R KADAM (182MCA54)

SHIFA SIDDIQUA (182MCA46)

Opinion Mining :

Opinions are statements that reflect people's perception or sentiment. These statements also provides opinion on objects or events. Opinion Mining or Sentiment analysis is a task under natural language processing for finding the mood of the customers about a purchasing of a particular product or topic. It involves building a system to collect and examine opinions about the product made in many online purchasing sites. Opinion mining is a sub field of web content mining. Web content mining is branch of Data mining. Opinion mining is a system which is used to identify and extract subjective information in text documents. Opinion mining is also called Sentimental analysis. The explosion of social media has created unprecedented opportunities for citizens to publicly voice their opinions, but has created serious bottlenecks when it comes to making sense of these opinions. At the same time, the urgency to gain a real-time understanding of citizens concerns has grown: because of the viral nature of social media (where attention is very unevenly and fast distributed) some issues rapidly and unpredictably become important through word-of-mouth. Policy-makers and citizens don't yet have an effective way to make sense of this mass conversation and interact meaningfully with thousands of others. As a result of this paradox, the public debate in social media is characterized by short-termismand auto-preferentiality. Many experts consider social media as a missed opportunity for better policy debate. At the same time, the sheer amount of raw data is also an opportunity to better make sense of opinions. The key asset that Google exploited to reach dominance in the search market is not a better algorithm, but the power of more data (quote).The research scope in opinion mining and sentiment analysis are:

1) Spam Detection using Sentiment Analysis. 2) Sentiment Analysis on short Sentence that include abbreviations.3) Improvement of existing sentiment word identification algorithm.4) Developing fully automatic tools for analysis.5) Effective Analysis of policy documents which containing opinion content.6) Managing the of bi polar sentiments successfully.7) Designing and Generation of highly content corpus database.

Sentiment Analysis:-

Sentiment analysis, opinion mining and subjectivity analysis are interrelated areas of research which use various techniques taken from Natural Language Processing (NLP), Information Retrieval (IR), structured and unstructured Data Mining (DM).

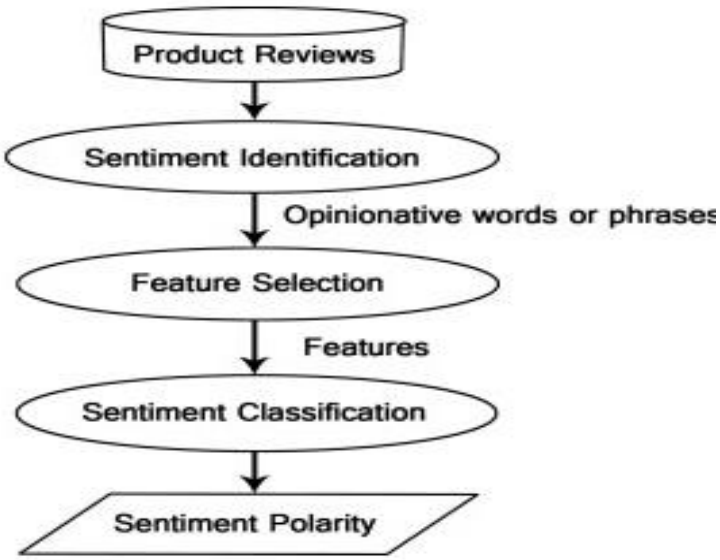
Sentiment analysis is not a single problem; instead it is a multi-faceted problem.There are three main classification levels in SA: document-level, sentence-level, and

aspect level SA.

a)**Document-level SA** : aims to classify an opinion document as expressing a positive or negative opinion or sentiment. It considers the whole document a basic information unit (talking about one topic). Sentence-level SA aims to classify sentiment expressed in each sentence.

b) **Sentence-level SA** : sentences are just short documents, the sentence level does not provide the necessary detail needed opinions on all aspects of the entity which is needed in many applications, to obtain these details; we need to go to the aspect level.

c) **Aspect-level SA** : aims to classify the sentiment with respect to the specific aspects of entities, to identify the entities and their aspects. The opinion holders can give different opinions for different aspects of the same entity like this sentence "The voice quality of this phone is not good, but the battery life is long".



Lexicon-based algorithms are frequently used to solve general SA problems because of their scalability.

They are also simple and computationally efficient shows the algorithms used. The data used in SA

are mostly on Product Reviews in the overall count.

References:

Isa Maks, Piek Vossen**A lexicon model for deep sentiment analysis and opinion mining applications,** Decis Support Syst, 53 (2012).

B. Pang, L. Lee **Opinion mining and sentiment analysis**
Found Trends Inform Retriev, 2 (2008).

A. Moreo, M. Romero, J.L. Castro, J.M. Zurita**Lexicon-based comments-oriented news sentiment analyzer system**
Expert Syst Appl, 39 (2012)

WEB MINING IN BIOINFORMATICS

ZAINAB FATHIMA(182MCA28)

GOWRI. A(182MCA25)

Introduction

Data Mining provides a way to gain particular knowledge from large number of databases. Data Mining in bioinformatics helps to find knowledge from different number of biomolecular data. BioInformatics combines the field of mathematics, engineering and statictics to find the pattern from biological data. It involves storing the data, extracting, analysis, prediction and usage of the information obtained from large biomolecular data. Examples are protein structure prediction, gene prediction, types cancer classification etc.

Databases involved in Bioinformatics

- **Medline**: It contains information about medicine, nursing, pharmacy, veterinary medicines etc.
- **Protein Databank**: Determining the protein structure by locating the atom in the molecules by X-ray or NMR scan.
- **Swiss-Prot**: It contains protein sequence stored in the database.
- **Nucleotide EMBL Database**: It stores the nucleotide sequence patterns as data in database.

Data mining tasks

- **Classification**: Arranges the data into their respective groups. Algorithms used: Naïve Bayes, Decision tree etc
- **Clustering:** groups the similar items together.
- **Association:** Based on the relationship between the variables it groups them.
- **Regression:** It finds the function to model the data.

Applications of data mining in bioinformatics

- **Genome Annotations** : It is used mark gene into DNA sequence.
- **Analysis of gene expression**: It takes the snapshot of protein structure and classifies them into groups.
- **Analysing mutations in cancer**: It uses an algorithm to compare the human genome sequencing results.
- **Protein Structure Prediction**: Amino acid sequence pattern of protein is stored and classified into different groups.
- **High throughput Image Analysis:** Clinical Visualized images are classified and information is extracted.

Following are the aspects in which data mining contributes for biological data analysis

1. Semantic integration of heterogeneous, distributed genomic and proteomic databases.
2. Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
3. Discovery of structural patterns and analysis of genetic networks and protein pathways.
4. Association and path analysis.
5. Visualization tools in genetic data analysis

Challenges Faced:

Data mining in bioinformatics is hampered by many facets of biological databases, including their size, number, diversity and the lack of a standard ontology to aid the querying of them as well as the heterogeneous data of the quality and provenance information they contain. Apart from these issues the following hindrances are faced

1. Which gene causes diseases? Finding out the kind of pattern in genes that can cause the disease is a challenging problem today for research by doctors and so providing the feature selection dataset to methodologies is crucial. But some learning mechanisms are adopted such as: manifold learning, semi-supervised learning.
2. Selection of data mining methods: Current data mining methods such as SVMs, discriminant analysis, neural networks give knowledge that is hard to understand by biologists
3. Can we build a unifying model for transferring the learned knowledge to classify/cluster other chromosomes(transfer learning) but old and new biological data may have different distributions
4. Exploiting the network structure: Real networks can be much more complex, involving thousands of genes, leading to the complex patterns of attractors and cell activities.

References:

1. https://www.researchgate.net/publication/323341042_A_Web_Mining_Application_to_Classify_Bioinformatics_Datasets
2. http://www.ijtrd.com/papers/IJTRD1357.pdf
3. https://arxiv.org/ftp/arxiv/papers/1205/1205.1125.pdf#:~:text=Applications%20of%20data%20mining%20to,cleansing%2C%20and%20protein%20sub%2Dcellular
4. https://www.bing.com/search?q=+bioinformatics&qs=n&form=QBRE&msbsrank=6_7&sp=-1&pq=bioinformatics&sc=6-14&sk=&cvid=B1C6A13EAEA1497CBC2E0802B7B2DE0E
5. https://www.bing.com/search?q=web+mining+in+bioinformatics&qs=n&form=QBRE&msbsrank=1_2&sp=-1&pq=web+mining+in+bioinformatics&sc=1-28&sk=&cvid=DD7784F8F79448A1AF6422E006863C78
6. https://ieeexplore.ieee.org/document/6122559
7. https://www.dbs.ifi.lmu.de/Forschung/Bioinformatics/
8. https://bioinformaticsonline.com/pages/view/918/data-mining-in-bioinformatics

ROLE OF WEB MINING IN E-COMMERCE

AKILANDESWARI T(18MCA02)

SHILPA K(18MCA17)

Introduction:

The main purpose of web mining is discovering information from the World-wide web and its usage patterns. E-commerce that is mainly characterized by doing business electronically with the help of internet had provided us a cost efficient and effective way of doing business. It is very important to apply web mining to e-commerce to gather information about user's and data.

Role of Web mining in E-commerce:

- **Financial Analysis:** It includes reviewing of revenues ,calculation, cost and comparative analysis of corporate income statements, cash flow statement, analysis of corporate balance sheet, analysis of financial markets and sophisticated controlling. Web mining be an effective tool.
- **Market Analysis:** It includes analysis of sales profitability, profit margins, meeting sales, sales receipt, time of orders, action undertaken by competitors, stock exchange quotations and market identification and segmentation. Web mining can be used here as a key tools that helps in building effective marketing strategy.
- **Logistic Analysis:** It can be effective to identify partners of supply chain quickly, reverse logistics analysis and handling.
- **Production Management analysis:** Where work is mainly to identify production 'bottlenecks' and delayed order and enabling organizations to examine production result obtained by plants or departments.
- **Customer Analysis:** It mainly concern time maintaining contacts with Customer profitability, modelling behaviour, customer and reaction customer satisfaction ,analysis etc. Web mining tells us what strategy should be used to get number of Customer with quality.
- **Web Analysis:** Where analysis of wage related data including wage components reports made with references to the type required, reports made from the respective of a given enterprise, wage report distinguishing employments type, personal contribution report, analysis of average wages.

Diagram for Web mining in commerce:

Conclusion:

Web mining is applied to e- commerce to know the browsing behaviour of customer ,to determine the success of marketing efforts, to improve the design of e-commerce website and to provide personalized services.

Reference

https://www.reserachgate.net/publication/272406319_Role_of_we_Mining_in_E-Commerce.

https://ieeexplore.ieee.org/document/5565964.

WEB MINING TECHNIQUES FOR EXTRACTION OF NEWS

SHRESTA R(182MCA34)
TEJASHWINI N(182MCA42)

Online news as an up-to-date and important information source, is an absorbing data repository for data mining. However, news content of most web pages is embedded in a large amount of noisy materials. Accurate extraction of news content is a necessary and crucial step for news text mining. One of the most important features is the similarity of the twin-pages which are collected from the same topic section of a site and published on the same/near date. A similarity measure based on edit distance is introduced and applied in the algorithms to separate the news content from noisy information. This method is much less complicated than other ones, and its accuracy and efficiency are fairly high, its complexity about the pages size is just linear.

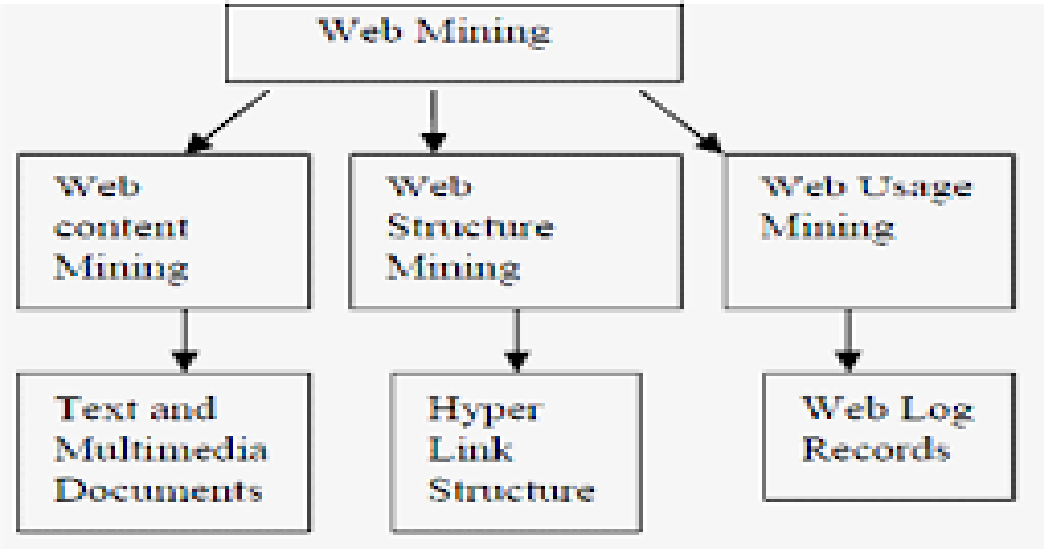
Web mining is the use of data mining techniques to automatically determine and abstract material from Web forms and facilities

Classification is a data mining method used to calculate group relationship for data requests. It is often denoted as supervised learning. It has a predefined traditional of groups or models built on that we forecast value. In this age of evidence, news is now simply available, as content suppliers and content radars such as online news facilities have developed on the World Wide Web. In the meantime the improvement of WWW, it is necessary to handle a very large quantity of automated data of which the majority is in the procedure of text. This situation can be successfully controlled by numerous Data Mining methods

Web newspapers provide a valuable resource for information. In order to benefit more from the available information, text mining techniques can be applied. However, because each newspaper page often covers a lot of unrelated topics, page-based data mining will not always give useful results. In order to improve on complete-page mining,

Web content mining uses different techniques

- Information Extraction
- Summarization
- Information Visualization
- Topic Tracking
- Categorization
- Clustering



Data mining is a concept that helps to find information which is needed from large data warehouses by using different techniques. It is also used to analyze past data and improve future strategies. Web data mining is considered as sub approach of data mining that focuses on gathering information from web. Web is a large domain that contains data in various forms i.e.: images, tables, text, videos, etc. As size of web is continuously increasing; it is becoming very challenging task to extract information. Each type has different algorithms, tools and techniques that are used for data retrieval. Various algorithms, tools and techniques for extraction. Web content mining is useful in terms of exploring data from text, table, images etc. Web structure mining classifies relationships between linked web pages. Web usage mining is also an important type that stores user access data and get information about specific user from logs.

References:

https://www.researchgate.net/publication/326180019 Data Mining: Web Data Mining Techniques, Tools and Algorithms

https://www.researchgate.net/publication/267948426 Extraction of News Content for Text Mining Based on Edit Distance

WEB MINING IN BUSINESS COMPUTING

NANCY CHELLAM E(182MCA45)
ISUKAPPALI DIVYA(182MCA39)

In the world of information technology(IT), everyone has the drift to do business electronically. Today lot of businesses are occurring on World Wide Web (WWW), it is very important key for the website owner to come up with a better platform to win over more customers for their site. Providing information in a fitter way is the solution to bring more customers or users. Customer is the end-user, who obtain the information in a way it submits some credit to the web site owners. In this article it defines web mining and present a method to make use of web mining in a better way to know the users and website etiquette which in turn enhance the web site facts to attract more users. This article also presents a web content mining.

Web mining is the subrange of data mining, which works with the removal of interesting knowledge from the WWW. Internet is a key for E-Business. E-Business is the electronic style of business which relies on internet. It is a medium for the retail to reach the customer and serve them in an improve way to make them to revisit their site. Sustaining one customer gives more customers to the particular vendor. Data mining is the understanding discovery technique from the huge amount of data. Enterprise supply planning online software is there to do outsourcing work. To mine the related user pattern from the huge pool of data, we can employ the web mining to improve the better understanding of the customer behaviour.

Today buying and selling are happening in web, even services also been done through internet. The interests of several research communities, the tremendous growth of information sources available on the web and recent interest in E-Business. Web mining field involves of main three categories, web usage mining, web structure mining and web content mining. In web usage mining the goal is to examine web page usage patterns in order to learn about a web system's users or the relationships between the forms. Web usage mining is useful for providing modified web services, an area of web mining research that has lately become active. To appeal the customers, it is very important for the web site owner to deliver the information in a different way from the other website.

Conclusion and Future work

Web mining methods have strong practical implication on e-system. Web data mining forms the basis of marketing and e-commerce activities on the web.it can also provide fast and efficient amenities to users as well as for businesses. Data mining in e-business will continue to be a very promising area of research.

References

- Advanced Computing: An International Journal (ACIJ), Vol.3, No.6, November 2012
- International Journal of Computer Applications (0975 – 8887) Volume 69– No.8, May 2013

WEB MINING FOR WEB PERSONALISATION

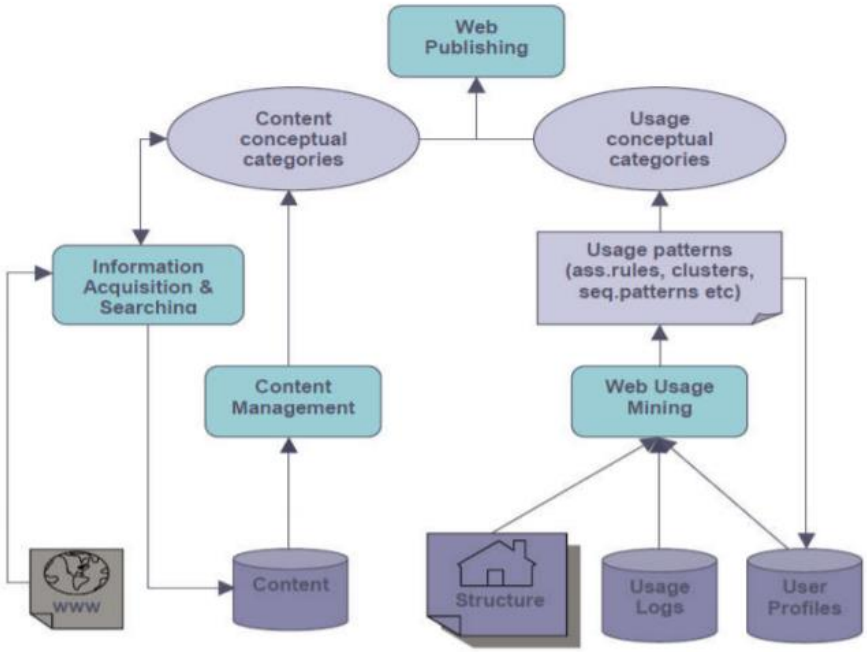
VIJAYALAKSHMI S (182MCA53)

POORNIMA S (182MCA38)

Introduction

Web personalization is the process of customizing a Web site to the needs of specific users, taking advantage of the knowledge acquired from the analysis of the user's navigational behaviour in correlation with other information collected in the Web context, namely, structure, content, and user profile data. Due to the explosive growth of the Web, the domain of Web personalization has gained great momentum both in the research and commercial areas, web usage mining is used to mining data from the web server log file.in web personalisation web usage mining deliver the particular content.

Modules of Web Personalisation system



Website Personalization is the process of creating customized experiences for visitors to a website. Rather than providing a single, broad experience, website personalization allows companies to present visitors with unique experiences tailored to their needs and desires.

Classified into different types

- **Content data** are represent to the end user, that can be simple, text, images or structured data such as structured data, such as information retrieved from databases.

- **Structure data** represent the way content is organized.it can be data entities used within web pages. Egs : HTML,XML
- **Usage data** represent a web site usage like IP address, time , date of access, complete file details.
- **User Profile data** provide information about the users of a website, it contains demographic information (name , age ,country ,education) for each user of a web site.
- **User Profiling** in the web domains, user profiling is the process of gathering information specific to each visitor, either explicitly, implicitly.

- **Log analysis and web usage mining**
  - Extract statistical information and discover interesting usage patterns.
  - Cluster the users into groups according to their navigational behaviour.
  - Discover potential correlations between web pages and user groups.
- **Content management process** of classifying the content of a web site in semantic categories in order to make information retrieval and presentation easier for the users. egs: news site or portals.
- **Web site publishing** a publishing mechanism is used in order to present the content stored locally in a web server and some information retrieved from other web resources in a uniform way to end-user different technologies can be used to publish data on the web.
- **Information acquisition and searching** searching and relevance ranking techniques must be employed both in the process of acquisition of relevant information and in publishing of the appropriate data to each group of users.
- **Content based filtering system** are solely based on individual user, it tracks each user behaviour and recommends items to user that user have liked it before.
- **Collaborative filtering system** invites users to rate objects and returns information that is predicted to be of interest to them.it based on the assumption of the users of similar behaviour.
- **Rule based filtering** the users are asked to take question or survey these are from the decision tree. after the results, it is trained according to the needs.
- **Log analysis and web usage mining**
  - Web usage mining is to reveal the knowledge hidden in the log files of the web server.by applying statistical and data mining methods to the web log data, interesting patterns concerning the users navigational behaviour can be identified, such as user and page clusters, as well possible correlation between web pages and user groups.
- **Web log access** to a web pages is recorded in the access log of the web server that hosts it.

  The entries of a web log file consist of fields that follows a predefined format.
- **Web data abstraction** concerning web usage, content and structure. the web characterization activity has published a drafts establishing precise semantics for concepts such as website, user, user session, servers, pageviews and clickstream.
-

- **Data preprocessing**
  - **Binning_Method:**
    This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.
  - **Regression:**
    Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).
  - **Clustering:**
    This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

Conclusion

Web personalization is the process of customizing the content and structure of a Web site to the specific and individual needs of each user, without requiring them to ask for it explicitly. This can be achieved by taking advantage of the user's navigational behaviour, as revealed through the processing of Web usage logs, as well as the user's characteristics and interests. Such information can be further analyzed in association with the content of a Web site, resulting in improvement of the system performance, users' retention, and/or site modification.

Reference:

1. **https://pdfs.semanticscholar.org/cc39/0c57b3816776240ed09169259b5c85ede55a.pdf?_ga=2.50588576.806550373.1603949891-727161119.1603949891**
2. **https://www.researchgate.net/publication/220169917_Web_mining_for_Web_personalization**
3. **https://link.springer.com/chapter/10.1007/978-3-540-72079-9_3**

SEMANTIC WEB MINING

JYOTSANA R JAIN(182MCA52)

PRANATHI S (182MCA37)

Semantic Web Mining aims at combining two fast-developing research areas namely the Semantic Web and Web Mining. The Semantic web and web mining are both built on the success of the world wide web (WWW).The Semantic Web is the second-generation WWW, enriched by machine-processable information supporting users in the tasks.

Researchers are exploiting semantic structures in the web to improve the results of web mining . They use web mining techniques to build the semantic web.These techniques are also used to mine the semantic web itself .Web Mining aims at discovering insights about the meaning of Web resources and their usage, formalizing the semantics of web sites and navigation behaviour.The nature of most data on the web is extremely unstructured. It could only be understood by humans, and the amount of data is so huge that it can only be processed efficiently by machines.

The Semantic Web addresses the first (initial) half of the challenge by trying to make the data (also) machine understandable, while Web Mining addresses the second half of the challenge by (semi-)automatically extracting the useful knowledge hidden in the data, making it available as an aggregation of manageable proportions.

The wording Semantic Web Mining emphasizes the spectrum of possible interactions between both research areas: it can be read as Semantic (Web Mining) or/and as (Semantic Web) Mining.The semantic information in text corpora is implicitly exploited by statistical methods in an attempt to “break the syntax barrier” on the Web analyzing the structural characteristics of data.

The backbone of the semantic web are the othologies ,which at present are hand-crafted and are not scalable solutions for wide range applications of semantic web technologies.It is challenging to learn ontologies, and/or instances of its concepts, in a (semi-)automatic way. For instance, search engines today are already quite powerful, but often return excessively large or inadequate lists of hits. Machine-processable information can point the search engine to the relevant pages and can thus improve both precision and recall.it is almost impossible to retrieve information with a keyword search when the information is spread over several pages.

Example:The query for Web mining experts in a company intranet, where the only explicit information stored are the relationships between people and the courses they attended on one hand, and between courses and the topics they cover on the other hand. In that case, the use of a rule stating that people who attended a course which was about a certain topic have knowledge about that topic might improve the results.

The steps show the direction where the Semantic Web is heading:

1. Providing a common syntax for machine understandable statements.
2. Establishing common vocabularies.
3. Agreeing on a logical language.
4. Using the language for exchanging proofs.

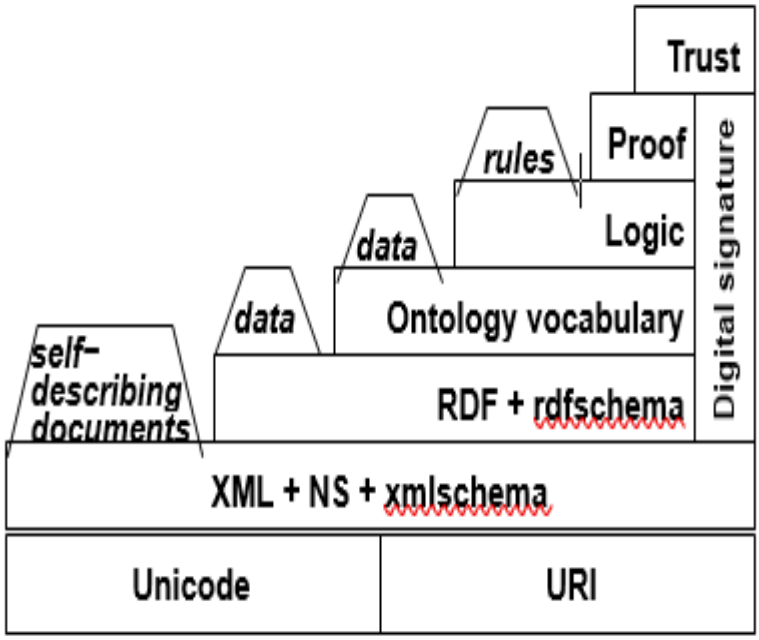


Fig - The layers of the Semantic Web

Web mining is distinguished in three areas namely content mining, structure mining, and usage mining in particular with association rule discovery, clustering, classification, and sequence mining.

Semantic web mining is an upcoming application that argues the two areas Web Mining and Semantic Web a need for each other to fulfill their goals to the full potential of their convergence to describe how two areas cooperate today and how the cooperation can be improved further.

References:

1. https://www.researchgate.net/publication/333641514_Semantic_Web_Mining_Survey_and_Analysis
2. https://www.researchgate.net/publication/292879064_A_Review_on_Semantic-Based_Web_Mining_and_its_Applications

STRUCTURED DATA EXTRACTION

AFRAH HASHMI A G(182MCA44)

AMTUL HASEEBA (182MCA49)

INTRODUCTION :

Structured data on the Web are typically data records retrieved from underlying databases and displayed in Web pages following some fixed templates. Extracting such data records is useful because it enables us to obtain and integrate data from multiple sources (Web sites and pages) to provide value-added services, e.g., customizable Web information gathering, comparative shopping, meta-search, etc. With more and more companies and organizations disseminating information on the Web, the ability to extract such data from Web pages is becoming increasingly important. At the time of writing this book, there are several companies working on extracting products sold online, product reviews, job postings, research publications, forum discussions, statistics data tables, news articles, search results, etc. Researchers and Internet companies started to work on the extraction problem from the middle of 1990s.

TYPES :

1. Manual approach: By observing a Web page and its source code, the human programmer finds some patterns and then writes a program to extract the target data. To make the process simpler for programmers, several pattern specification languages and user interfaces have been built. However, this approach is not scalable to a large number of sites.

2. Wrapper induction: This is the supervised learning approach, and is semi-automatic. The work started around 1995-1996. In this approach, a set of extraction rules is learned from a collection of manually labeled pages or data records. The rules are then employed to extract target data items from other similarly formatted pages.

3. Automatic extraction: This is the unsupervised approach started around 1998. Given a single or multiple pages, it automatically finds patterns or grammars from them for data extraction. Since this approach eliminates the manual labeling effort, it can scale up data extraction to a huge number of sites and pages.

USE OF STRUCTURED DATA EXTRACTION :

Extracting structured data from the web pages is clearly very useful, since it enables us to pose complex queries over the data. Extracting structured data has also been recognized as an important sub-problem in information integration systems , which integrate the data present in different web-sites. Therefore, there has been a lot of recent research in the database and AI communities on the problem of extracting data from web pages (sometimes called information extraction (IE) problem).

REFERENCES :

- https://link.springer.com/chapter/10.1007/978-3-642-19460-3_9
- https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/extract.pdf