JYOTI NIVAS COLLEGE
AUTONOMOUS

EJOURNAL: DEPARTMENT OF MCA

ISSUE: 4 **OCTOBER 2018**

INDEX

SL NO	TITLE	PG NO
1	Web mining in e-commerce	1
2	Managing website structure information in web mining	3
3	Clustering with efficient web usage mining	5
4	Web crime mining by means of data mining techniques	6
5	Block chain in web data mining	8
6	Extracting and analysing web social networks	10
7	Weather Forecasting using Data Mining	11
8	Web structure mining	14
9	Opinion mining and sentimental analysis using web mining	15
10	Architectural pattern mining	17
11	Web semantics	18
12	Using data mining to make sense of climate change	20
13	Web information retrieval	21
14	Web personalization on web usage mining	22
15	E-learning in web mining	23

WEB MINING IN E-COMMERCE

AKILA Y R (16MCA01)
B T VEENA (16MCA06)

Introduction

The World Wide Web is the key source of information and it is growing rapidly and has become much accepted over the last decade, bringing a strong platform for information distribution, retrieval and analysis of information. One of the techniques used in World Wide Web for data retrieval to discover useful information is Web Mining.

Web mining is applied to extract the interesting, useful patterns and hidden information from the Web documents and Web activities such as Web pages, media objects on the Web, Web links, Web log data, and other data generated by the usage of Web data. Web mining extends analysis much further by combining other corporate information with Web traffic data. Practical applications of Web mining technology are abundant and are by no means the limit to this technology. E-commerce is one of the main applications of web mining.

E-commerce is all about carrying out business on the Web. It is about carrying out transactions, essentially buying and selling products and services by consumers and businesses on the web. E-commerce websites have the advantage of reaching many customers regardless of distance and time limitations. The advantage of it over traditional businesses is the faster speed and the lower expenses for both e-commerce website owners and customers in completing customer transactions and orders. It not only keeps the business up and running but also provides a cost efficient and effective way of doing business in the web.

Web Mining Tools in E-commerce

Different data mining tools are used depending on different mining goals, there are major three categories.

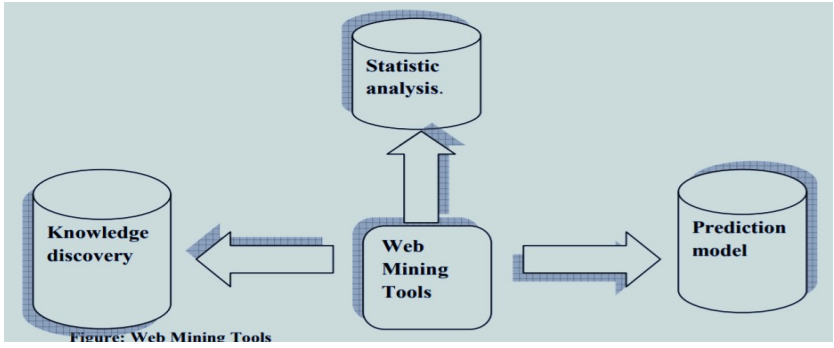
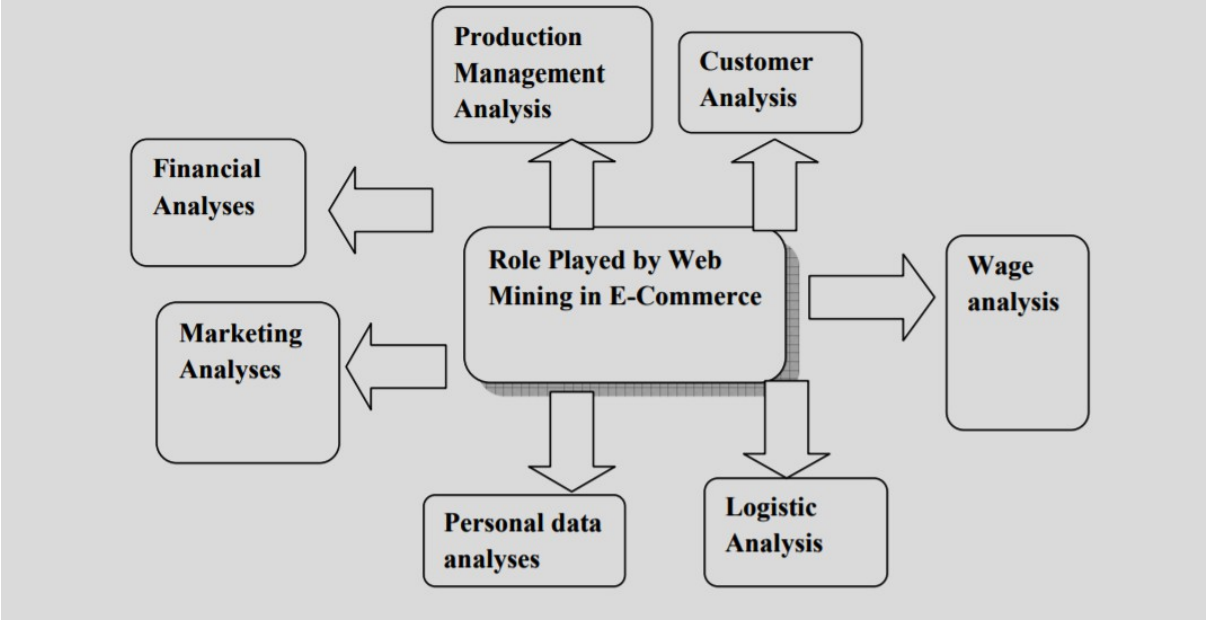


Figure: Web Mining Tools

1) **Statistic analysis:** This method is basically used to check the math rules in data and utilize statistic modes and math models to interpret these rules. This method helps to find the identification of time series data patterns and anomalies in the data.
2) **Knowledge discovery:** This is obtained from artificial intelligence and machine learning, which uses a data-search process, to extract information from the data, as well as the relationship between data elements and models from which to discover business rules and business facts.

3) **Prediction model**: This model is based on consumer behaviour has a certain repetitive and regularity of such a hypothesis, which allows businesses to collect stored in the database b analysing the transaction information to predict consumer behaviour.

Roles Played by Web Mining in E-Commerce



Conclusion

The growth of World Wide Web and technologies has made business functions to be executed fast and easier. As large amount of transactions is performed through e-commerce sites and the huge amount of data is stored, valuable knowledge can be obtained by applying the Web Mining techniques. Using Web Mining, companies can understand customer behaviour, improve customer services and relationship and measure the success of marketing efforts.

MANAGING WEBSITE STRUCTURE INFORMATION IN WEB MINING

MARIA MARGARET J
(162MCA31)
SHYLA J(162MCA33)

Introduction

Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web. (Mining means extracting something useful or valuable from a baser substance, such as mining gold from the earth.) Web mining is used to understand customer behaviour, evaluate the effectiveness of a particular Web site, and help quantify the success of a marketing campaign.
The significant web mining applications are website design, web search, search engines, information retrieval, network management, Ecommerce, business and artificial intelligence, web market places and web communities. Online business breaks the barrier of time and space as compared to the physical office business. Big companies around the world are realizing that e-commerce is not just buying and selling over Internet, rather it improves the efficiency to compete with other giants in the market. This application includes the temporal issues for the users. Challenges in Web Mining is information can be huge, redundant and diverse.

General elements of website Structure management

- **Clean up Any Post Launch Tasks:** Go through all posts, images, screen size, signup and pathways.
- **Create and Execute Marketing Strategies:** This involves social media, content marketing, and email marketing, search engine optimization, pay per click.
- **Continuous Improvement and Optimization:** Continuous check on software's to check headlines, media and display social proof.
- **General Website Maintenance and Protection:** Should have regular website backup, ensure themes and plugins are up to date and do a routine website text.

Web structure mining uses graph theory to analyse the node and connection structure of a web site. Web structure mining can be divided into two kinds:

1. Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location.
2. Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

Features and Benefits

1. Information filtering techniques try to learn about user's interests based on their evaluation and actions and then to use this information to analyse ne documents.
2. It increase the value of each visitor and improve the visitor's experience at the websites.
3. It allows you to look for pattern in data through content mining, structure mining and usage mining.

CLUSTERING WITH EFFICIENT WEB USAGE MINING

INDUMATHI S (16MCA09)
AMRUTHA A (162MCA26)

Introduction

Web mining is the application of data mining techniques to extract knowledge from web data, including web documents, hyperlinks between documents, usage logs of web sites, etc. Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to web content has been the most widely researched.

Clustering with web mining has drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision, the application of these techniques to web content mining has been limited.

Web usage mining attempts to discover useful knowledge the secondary data obtained from the interactions of the users with the Web. This process involves two Algorithms. A hybrid evolutionary **Fuzzy clustering algorithm** is proposed to optimally segregate similar user interests. FCM algorithm provides an iterative approach to approximate the minimum of the objective function starting from a given position and leads to any of its local minima. No guarantee ensures that FCM converges to an optimum solution. The algorithm is initialized by constraining the initial values to be within the space defined by the vectors to be clustered.

Expectation maximization (EM) is used for clustering in the context of mixture models. This method estimates missing parameters of probabilistic models. Generally, this is an optimization approach, which had given some initial approximation of the cluster parameters, iteratively performs two steps, i.e., the expectation step computes the values expected for the cluster probabilities, and second, the maximization step computes the distribution parameters and their likelihood given the data. It iterates until the parameters being optimized reach a fix point or until the log-likelihood function, which measures the quality of clustering, reaches its maximum.

The algorithm is similar to the Fuzzy C-means procedure in that a set of parameters are re-computed until a desired convergence value is achieved.

WEB CRIME MINING BY MEANS OF DATA MINING TECHNIQUES

PREMA D (16MCA14)
VINISHA R (16MCA25)

Introduction

Criminal web data always offer valuable and appropriate information for Law administration. The evaluation of the different capacities of wide spread criminal web data is very difficult all the time so it is one of the most noteworthy tasks for law administration. Crimes may be as extreme as murder and rape where advanced analytical methods are required to extract useful information from the data Web mining comes in as a solution.
Definitely, one of influential factors that encounter a crime phenomenon is the humans' social life circumstances so the crime analysis knowledge is needed as an efficient combating tool. It also comprises of leveraging a systematic approach for discovering, identifying and predicting crime incidents and its input is contained assigned information and data in crime variables and the output contains the answer to knowledge extraction, analytical and investigative questions and the visualization of the results. The aim of web mining is to extract appropriate information from the page content, Web hyperlink structure and usage data. Although Web mining uses many data mining techniques, it is not purely an application of traditional data mining due to the heterogeneity and semi-structured or unstructured nature of the Web data.

Web Crime Mining

All intelligence-gathering and law-enforcement organizations major challenge is facing to the efficient and correct evaluating of the crime data growing volumes. One of the examples of this can be complex conspiracies that are often hard to undo since the knowledge of suspects can be geographically span and diffuse in the long time. Detecting cybercrime can be very hard as well, because of frequent online transactions and busy network traffic which create huge amounts of data and just a portion of which relates to illegal activities.

- Facing to the huge amount of information on the Web that is very wide and diverse so any user can find information on almost anything on the Web.
- Huge amount of data from all types are exist in unstructured texts, semi-structured Web pages structured tables and multimedia files.

- The information on the Web is noisy that is comes from two main sources. The first one is that a typical Web page involves many pieces of information for instance the navigation links, main content of the page, copyright notices, advertisements and privacy policies. Only part of the information is useful for a particular application but the rest is considered noise. For performing a fine-grain, the data mining and Web information analysis, the noise should be removed. The second one is due to the fact that the Web does not have quality control of information, for example, a large amount of information on the Web is of low quality because any one can write everything
- The Web is about services for example most commercial Web sites allow the users to perform useful operations at their sites such as paying bills, purchasing products and filling the forms.

Crime Data Mining Techniques

The traditional data mining techniques just classify the patterns in structured data for example, classification and prediction, association analysis, outlier analysis and cluster analysis. On the other hand, the newer techniques identify patterns from unstructured and structured data. Crime data mining increases the privacy concerns like the other forms of data mining. However, the researchers' effort to promote the various automated data mining techniques for national security applications and local law enforcement. Particular patterns are identifies by Entity extraction from data such as images, text, or audio materials that has been utilized to automatically identify addresses, persons, vehicles and personal characteristics from police narrative reports. In computer forensics, the extraction of software metrics which includes the data structure, program flow, organization and quantity of comments and use of variable name scan facilitate further investigation by, for example, grouping similar programs written by hackers and tracing their behaviour. Entity extraction provides basic information for crime analysis, but its performance depends greatly on the availability of extensive amounts of clean input data.

The main techniques of the crime data mining are clustering, association rule mining, classification and sequential pattern mining. Although all of these efforts, the crime Web mining still is a highly complex task.

BLOCK CHAIN IN WEB DATA MINING

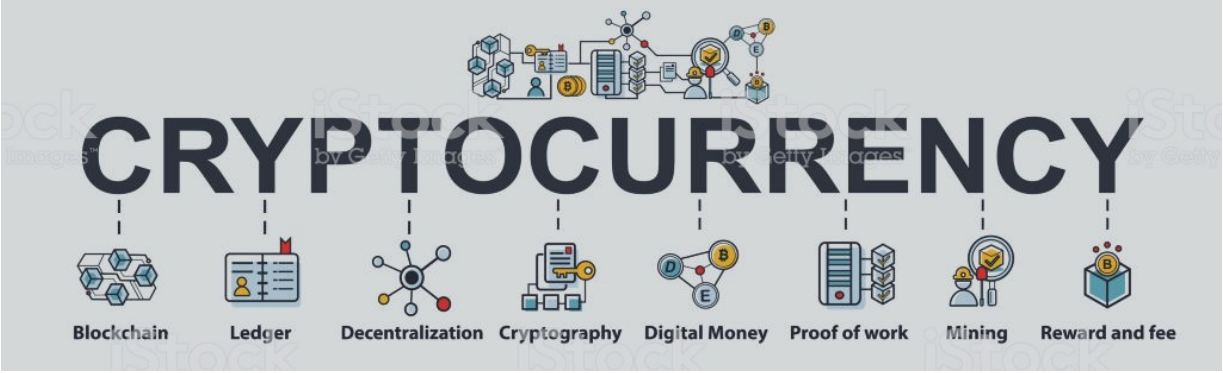
DIVYA MAJALIKAR (16MCA10)

SUPRIYA SUDHIR (16MCA21)

Introduction

Crypto currency mining is a process in which transactions are verified and added to block chain. This process is also known as Crypto Mining. Crypto currency mining has increased its usage and grown exponentially in last few years.

Each time a transaction is made, the crypto currency miner has to authenticate the information and update the block chain with the transaction. The mining process competes with other crypto miners to solve complex mathematical problems with cryptographic hash that are associated with the transaction of data.



There are quite web limitations that lead to data leak and data breaches:

- **Web 2.0 Limitations**

1) Data Intermediaries

→Once the data is out there, it's no longer yours. Some companies still make you feel that you have your own data but the reality is they have your data stored on their servers.

2) High Data Vulnerability

→There have been many data leaks over past decades. But since we have come across new hacking technology and poor data host security we have at least overcome some issues suffered by people. Over 500 companies like eBay, Yahoo, Uber and many more have suffered data breaches.

3) Data Trading

→The recent Cambridge Analytical/Facebook scandal is the best reminder that how data trading can be done so easily and threat fully.

→The amount of data that we handover to marketers and advertisers, what if the data has been reached in the people of wrong hands?

4) Need for trust

→Whenever you give your data to a company in exchange to use their services , you mostly trust the company that they will keep your data safe, but the problem is we don't even know , what they intend to do with our data.

→Basically, there should be 'no need for trust', instead it's your responsibility that when you give your data to a company, and you should know how to keep your data secure before you give it to a company. As we have discussed in the above points, block chain will help the transition of our web mining.

- **How web 3.0 projects will fix these issues?**

→Data mining in few companies have dominated the internet and gained access to 80 percent of all the data. That's how the internet was not provided free.

→Then the invention came of internet censorship bills and needed to be truly decentralized and anonymous.

→Everything you do online can be tracked.

Ex: Google knows where you are at based on their location maps. This identifies that in the upcoming days no one will truly have an own private life using internet. That is why we need web 3.0.

- **Promising web 3.0 projects**

There are already Projects in different sectors that provide some solutions and also help the internet to evolve from its present state. Most of the people will not be able to see the transition, but when it does, the world will be a better place. Most of the projects will be cheaper and more efficient to certain problems.

Ex: Substratum.

As we have discussed about few promising block chain projects that will help transition our web from 2.0 to 3.0. It will be possible that these projects will take some time before we see some good results and these problem can be solved when the developers will completely understand the block chain technology. And once it is done, the internet will be totally different place and experience.

EXTRACTING AND ANALYSING WEB SOCIAL NETWORKS

SOUMYA SUMAN(16MCA20)
SHEEBA KAUSAR(16MCA17)

Introduction

Social network analysis seems to be a traditional topic in the areas of psychology, sociology and behaviour science, it is becoming an active and popular topic in computer science domain due to its interdisciplinary research essence recently, especially with the propagation of Web 2.0 technology. Web communities, and capturing the dynamic evolution patterns of Web networked structures, and an empirical study of identifying social sense from a large scale Web log archive.

A Web community is a collection of Web pages created by individuals or associations with a common interest on a topic, such as fan pages of a baseball team, and official pages of computer vendors. Recent research on link analysis [52, 55, 70, 93, 99, 137,153] shows that we can identify a Web community on a topic by extracting densely connected structure in the

Web graph, in which nodes are Web pages and edges are hyperlinks. The Web community slightly differs from a community of people, for example, a Web community may include competing companies. There are several algorithms for finding Web communities.
Here, the extraction of Web community utilizes Web community chart that is a graph of communities, in which related communities are connected by weighted edges. The main advantage of the Web community chart is existence of relevance between communities. We can navigate through related communities, and locate evolution around a particular community.
M.Toyoda and M. Kitsuregawa explain how Web communities evolve, and what kinds of metrics can measure degree of the evolution, such as growth rate and novelty. They first explain the details of changes of Web communities, and then introduce evolution metrics that can be used for finding patterns of evolution. Here the notations used are summarized in this section.
t1,t2,….,tn:Time when each archive crawled. Currently, a month is used as the unit time.
W(tk): The Web archive at time tk.

C(tk): The Web community chart at time tk.

C(tk),d(tk),e(tk),….Communities in C(tk).

Types of Changes:

Emerge: A community c(tk) emerges in (tk), when c(tk) shares no URLs with any community in C(tk−1). Note that not all URLs in c(tk)newly appear in W(tk). Some URLs in c(tk) may be included in W(tk−1), and do not have enough connectivity to form a community

Dissolve: A community c(tk−1) in C(tk1) has dissolved, when c(tk−1)shares no URLs with any community in C(tk). Note that not all URLs in c(tk−1) disappeared from W(tk−1). Some URLs in c(tk-1) may still be included in W(tk) losing connectivity to any community.

WEATHER FORECASTING USING DATA MINING

NAVAMI V (162MCA32)
LATA SUDI

(162MCA30)

Introduction

Weather Prediction is the application of science and technology to predict atmospheric conditions ahead of time for a particular region. Prediction is one of the basic goals of Data Mining. Data Mining is to dig our knowledge and rules, which are hidden and unknown. User may be interested in or has potential value for decision-making from the large amounts of data. Such potential knowledge and rules can reveal the laws between the data. There are many kinds of technical methods of data mining, which mainly include: association rule mining algorithm, decision tree classification algorithm, clustering algorithm and time series

mining algorithm, etc. How to store, manage and use these massive meteorological data, discover and understand the law and knowledge of the data, to contribute to weather forecasting completely and effectively has attracted more and more data mining researcher's attention. This article construct the weather forecasting platform, using data mining for meteorological forecast and forecasts results are analysed.

Data mining algorithm

Data Mining Algorithms Five data-mining algorithms, neural network (NN), random forest, classification and regression tree (C&RT), support vector machine (SVM), and k-nearest neighbour (k-NN) were used to build the prediction models. NN consists of a group of interconnected neurons, making it an adaptive system that can change its structure based on external or internal information flowing through the network during the learning phase. NNs are usually used to model complex relationships between input and output variables. Random forest combines decision tree predictors in a way that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. It integrates a bagging idea and a random selection of features in constructing a collection of decision trees. C&RT, popularized by Breiman, is a nonparametric technique producing logical if–then rules that are easy to interpret. An SVM is a supervised learning method used for classification and regression analysis. SVM constructs one or a set of hyper planes in a high or infinite dimensional space. The key advantage of SVM is the use of kernel functions making SVM suitable for modelling in complex nonlinear domains. K-NN is an instance-based learning method accounting for contributions of the neighbours. It offers good performance for some classes of applications.

Data and methodology

eather forecasting completely and effectively has attracted more and more Data Mining researcher's attention[2]. This article constructs the Weather Forecasting platform, using data mining for

weather forecasting completely and effectively has attracted more and more Data Mining researcher'attention[2]. This article constructs the Weather Forecasting platform, using data mining for meteorological forecast and the forecast results are analyzed.

1. **Data collection**

Data collection is the process gathering of the required data for the processing purpose. We have collected from the locale city centre and crawled from the open weather site for the particular region. The data has been collected of the last few years. We have taken few parameters such as temperature, wind humidity and weather conditions.

2. **Data Cleaning**

In this stage, a consistent format for the data model was developed which took care of missing data, finding duplicated data, and weeding out of bad data. Finally, the cleaned data were transformed into a format suitable for data mining.

3. **Data Selection**

At this stage, data relevant to the analysis was decided on and retrieved from the dataset. The meteorological dataset had eight (8) attributes, their type and description is presented in Table 1, while an analysis of the numeric values are presented in Table 2. Due to the nature of the Cloud Form data where all the values are the same and the high percentage of missing values in the sunshine data both were not used in the analysis.

Table 4.1.2: Attributes of Meteorological Dataset

Attribute	Type	Description
Month	Numerical	Month consider
Year	Numerical	Year consider
Temperature	Numerical	Monthly min temp
Humidity	Numerical	Monthly min hum
Wind speed	Numerical	Wind run in km
Rainfall	Numerical	Total monthly rainfall

4. **Data Transformation**

This is also known as data consolidation. It is the stage in which the selected data is transformed into forms appropriate for data mining. The data file was saved in Commas Separated Value (CVS) file format and the datasets were normalized to reduce the effect of scaling on the data. Analyses of numeric values are shown in following tabl**e:**

No	variable	Min	Max	Mean	MAE	MSE	SD
1	Month	June(1)	June(30)	-	-	-	-
2	Year	2015	2015	-	-	-	-
3	Temperature	27	32	29.5	0.5957	1.0638	0.6981
4	Humidity	73	82	77.5	0.5957	1.0638	0.6981
5	Wind speed	6	23	14.5	0.5957	1.0638	0.6981
6	Rainfall	Rain	Rain	Rain	Rain	Rain	Rain

5. **Data Mining Stage**

The data mining stage was divided into three phases. At each phase all the algorithms were used to analyse the meteorological datasets. The testing method adopted for this research was percentage split that train on a percentage of the dataset, cross validate on it and test on the remaining percentage. Thereafter interesting patterns representing knowledge were identified.

Weather forecasting completely and effectively has attracted more and more Data Mining researcher's attention [2]. This article constructs the Weather Forecasting platform, using data mining for meteorological forecast and the forecast results are analysed.

WEB STRUCTURE MINING

SINDHU PRIYA (16MCA19)
SHIPRA KUMARI (16MCA18)

Introduction

The goal of web structure mining is to generate structural summary about web pages and web sites. It shows the relationship between the user and the web. It discovers the link structure of hyperlinks at the inter document level. Two algorithms that have been proposed to lead with those potential correlations: HITS and PageRank. In recent days the data generation is enormous in all the fields. Same as in Internet the data generation is high and there is no control over the data generation. To retrieve the exact data required by the online consumer is a tedious task. To achieve the same is done by data mining methods and its techniques. The data mining concept consist of web mining methods. The term web mining extracts the required information to user and to reach the necessary goal in the website. To attain the goal, use the concept of web mining. Web mining divides into web content, web structure and usage mining. Web structure mining plays very significant role in web mining process. The future algorithms for web structure mining such as Pagerank Algorithm, HITS, Weighted Pagerank Algorithm, Weighted page content rank Algorithm (WPCR) and soon. In this paper, identify their strengths and limitations of different algorithms used in web mining.

The WWW is one of the most important resources for information generation and the retrievals of data also an eminent step in web. The knowledge is discovered with the help of stable increasing of the amount of data generated in online. Considering the web aspect, the online users get easily lost in the web's loaded hyper structure. Through the available application of data mining methods leads to the perfect solution for knowledge discovery on the Web. The knowledge extracted from the Web can be used to raise the performances for Web information retrievals, question answering and Web based data warehousing. Web structure mining, one of three categories of web mining for data, is a tool used to identify the relationship between Web pages linked by information or direct link connection. It offers information about how different pages are linked together to form this huge web. Web Structure Mining finds hidden basic structures and uses hyperlinks for more web applications such as web search.

Web structure mining is the process of analyzing the hyperlink and mine important information from it and steps to achieve the information is tedious one. The primary objective of the Web Structure Mining is to generate the structural synopsis about the Web site and Web page. Web Structure mining will sort out the Web pages in different category and from the category to generate the information like the similarity and relationship between different Web sites.

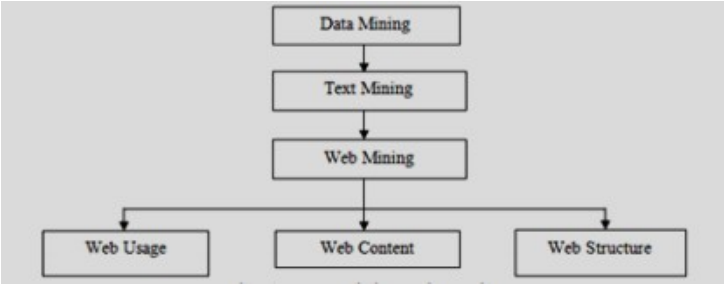
OPINION MINING AND SENTIMENTAL ANALYSIS USING WEB MINING

ANARKHA M (16MCA02)
GILDA RACHEL FERNANDEZ (162MCA27)

Introduction

Web mining is an area of sub discipline from text mining which aims in mining the semi structured data in the form of Web content mining, Web structure mining and Web usage mining.

Sentimental analysis also known as Opinion mining is used in analysing the important opinion from the reviews generated by the users. When any decisions are to be made regarding the purchase of new product, software or electronic products, people are very much interested in obtaining the reviews from various websites, blogs or discussion forums. In such case opinion mining or sentimental analysis is used widely which deals with tracking the mood of the people regarding a particular product or topic.



The topic can be an event, movie, location, drug, product, hotel etc. It is a Natural Language Processing and Information Extraction that aims in obtaining the feelings of the writer that are given by the positive or negative comments, by analysing a huge volume of documents.

Techniques in opinion mining

The data mining algorithms can be classified into different types of approaches as Supervised, Unsupervised or Semi - supervised algorithms.

A. Classification: Classification is the Supervised technique in which every instances belongs to the specific class, it is being indicated by the value of class attribute. Main goal of the classification algorithm is to improve the predictive accuracy in training the model. The algorithms being discussed include the following:

- K-Nearest Neighbour,
- Support Vector Machines.

1) **K-Nearest Neighbour:** K-Nearest Neighbour algorithm is being widely used for classification and regression and also it is a non-parametric method.

2) Support Vector Machines: SVM is widely being used for classification, regression and pattern recognition. SVM has capability to classify indeed of the dimensions or size of the input space.

B. Clustering: Clustering is the unsupervised technique that performs natural grouping of instances. It is the method of dividing the data into different groups with the similar objects. An effective clustering algorithm aims in obtaining the effective clusters irrespective of their shapes and size of data. Most Commonly used algorithms:

- K-Means Clustering

- SOM (self-organized map)

1) K-means clustering algorithm: K-means algorithm is most common and popular clustering tool that is widely used in many applications and it falls under the partitioning algorithms that aims in constructing the various patterns and evaluates them by using some criterion.

2) Self organized Map (SOM) algorithm: SOM is a type of the artificial neural network (ANN) that is unsupervised learning methodology. It is widely used in vector quantization and it belongs to the category of competitive learning networks.

Challenges in opinion mining:

- Apart from the noun words, Adjectives and verbs are also considered as feature words in some cases and it becomes difficult to classify.
- Customer can use abbreviations, short words or roman letters. For example lex for lexical, cam for camera etc., so it takes time to deal with such type of words and understanding it for the mining process.
- For The user opinions about various products, feedback would be on different language (French, Chinese, and Greek), so it becomes difficult to tackle each language with its orientation is difficult task and challenging.
- Web contains the various spam contents and it becomes difficult in eliminating the spam and fake reviews before processing to obtain the better accuracy in results.

Conclusions

We introduced and surveyed the field of sentiment analysis and opinion mining. It tried to showcase from basic definitions, different techniques, various evaluation methods that are commonly used for Sentiment Analysis. For obtaining the solution to any type of problems, dataset becomes the key factor and once dataset is chosen any kind of mining algorithms can be explored. Some of the algorithms that are most widely used in sentimental analysis and opinion mining are Support Vector Machine-Nearest Neighbour, K-means Clustering and SOM (Artificial Neural network).

ARCHITECTURAL PATTERN MINING

JANISHA K P (162MCA28)

SOUNDARYA R (162MCA30)

Introduction

Knowledge about the reception of architectural structures is crucial for architects and urban planners. Yet obtaining such information has been a challenging and costly activity. However, with the advent of the Web, a vast amount of structured and unstructured data describing architectural structures has become available publicly. This includes information about the perception and use of buildings (for instance, through social media), and structured information about the building's features and characteristics (for instance, through public Linked Data). Hence, first mining (i) the popularity of buildings from the social Web and (ii) then correlating such rankings with certain features of buildings, can provide an efficient method to identify successful architectural patterns. In this paper we propose an approach to rank buildings through the automated mining of Flicker meta-data. By further correlating such rankings with building properties described in Linked Data we are able to identify popular patterns for particular building types (airports, bridges, churches, halls, and skyscrapers). The approach combines crowd sourcing with Web mining techniques to establish influential factors, as well as ground truth to evaluate our rankings.

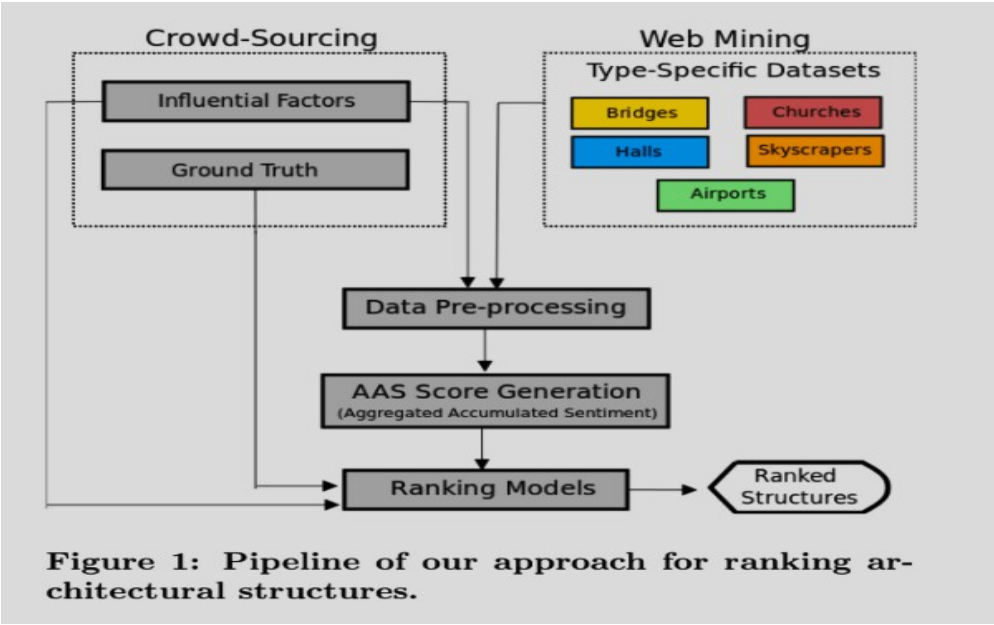


Figure 1: Pipeline of our approach for ranking architectural structures.

WEB SEMANTICS

ARCHANA P (16MCA04)
HEMA MALINI R (16MCA08)

Introduction

The semantic web is similar to the World Wide Web, created by Sir Tim Berners-Lee in 1989. However, rather than focusing on documents, it's instead built upon data. The W3C's official definition of the semantic web is "a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. The semantic web essentially allows for the connection of information using a network that can be easily read by machines – whether computers, IOT devices, mobile phones or other devices commonly used to access information.

The word semantic itself implies meaning or understanding. As such, the fundamental difference between Semantic Web technologies and other technologies related to data is that the Semantic Web is concerned with the meaning and not the structure of data. Note: Other semantic technologies include Natural Language Processing (NLP) and Semantic Search.

This fundamental difference engenders a completely different outlook on how storing, querying, and displaying information might be approached. Some applications, such as those that refer to a large amount of data from many different sources, benefit enormously from this feature.

From a technical point of view, the Semantic Web consists primarily of three technical standards:

- RDF (Resource Description Framework): The data modelling language for the Semantic Web. All Semantic Web information is stored and represented in the RDF.
- SPARQL (SPARQL Protocol and RDF Query Language): The query language of the Semantic Web. It is specifically designed to query data across various systems.
- OWL (Web Ontology Language): The schema language, or knowledge representation (KR) language, of the Semantic Web. OWL enables you to define concepts composably so that these concepts can be reused as much and as often as possible. Composability means that each concept is carefully defined so that it can be selected and assembled in various combinations with other concepts as needed for many different applications and purposes.

Benefits

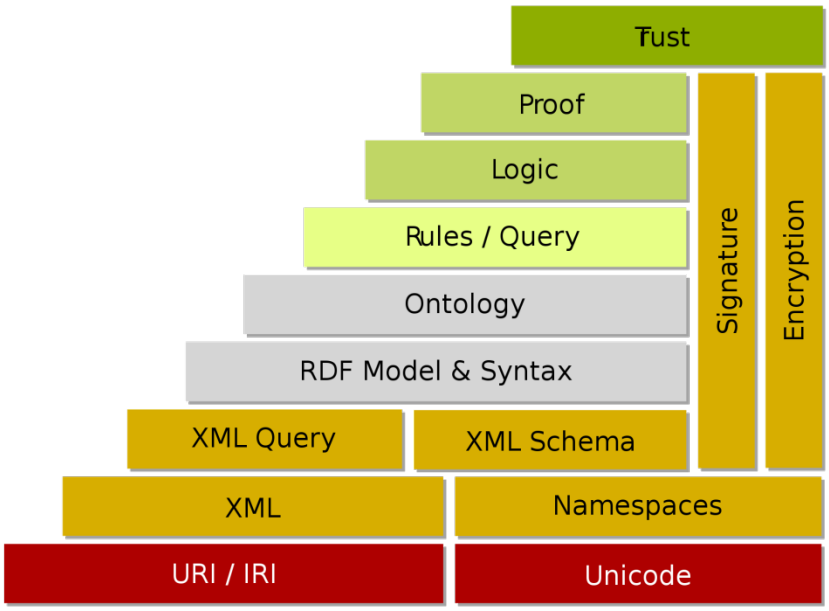
- One of the key benefits of the semantic web is having large amounts of data, knowledge and information made understandable and accessible to machines, especially artificially intelligent bots, virtual assistants and agents.
- The simplicity of the RDF data structure and the schema’s optional nature mean that it’s easy to combine different sets of data. This is particularly useful for big data projects where the variety of data within an organisation can present a challenge.

Conclusion

Right now the semantic web techniques cannot replace a human. He still must validate all the results that a computer generates. Still the human is the one to formal define concepts, things, and events, real live and presented in a machine-understandable form.

Even if the vision about the Web of trust can be still far way, we have to point out the important steps already achieved: RDF and OWL standards have been completed; many semantic web applications have been developed in the last years making collaboration with corporations much stronger, and given the right benefits to semantic web technologies.
There is much to be fulfilled but the opportunities are big because of the incredible capacity humans have, called knowledge

W3C Semantic Web Layers



USING DATA MINING TO MAKE SENSE OF CLIMATE CHANGE

ATHULYA U K (16MCA05)
PAVITHRA S (16MCA13)

Introduction

Big data and data mining have provided several breakthroughs in fields such as health informatics, smart cities and marketing. The same techniques, however, have not delivered consistent key findings for climate change.

"It's not that simple in climate," said Annalisa Bracco, a professor in Georgia Tech's School of Earth and Atmospheric Sciences. "Even weak connections between very different regions on the globe may result from an underlying physical phenomenon. Imposing thresholds and throwing out weak connections would halt everything. Instead, a climate scientist's expertise is the key step to finding commonalities across very different data sets or fields to explore how robust they are."

And with millions of data points spread out around the globe, Bracco said current models rely too much on human expertise to make sense of the output. She and her colleagues wanted to develop a methodology that depends more on actual data rather than a researcher's interpretation.

Tech team has developed a new way of mining data from climate data sets that is more self-contained than traditional tools. The methodology brings out commonalities of data sets without as much expertise from the user, allowing scientists to trust the data and get more robust -- and transparent -- results.

The methodology is open source and currently available to scientists around the world. The Georgia Tech researchers are already using it to explore sea surface temperature and cloud field data, two aspects that profoundly affect the planet's climate.

"There are so many factors -- cloud data, aerosols and wind fields, for example -- that interact to generate climate and drive climate change,"

"The methodology reduces the complexity of millions of data points to the bare essentials -- sometimes as few as 10 regions that interact with each other," said Nenes. "We need to have tools that reduce the complexity of model output to understand them better and evaluate if they are providing the correct results for the right reasons."

"Climate science is a 'data-heavy' discipline with many intellectually interesting questions that can benefit from computational modeling and prediction," said Dovrolis, a professor in the School of Computer Science, "Cross-disciplinary collaborations are challenging at first -- every discipline has its own language, preferred approach and research culture -- but they can be quite rewarding at the end."

WEB INFORMATION RETRIEVAL

SAYANA P (16 MCA 16)
MICHAEL SOPHIA A (16 MCA 12)

Introduction

Information Retrieval (IR) is dealing with the storage, representation and management of information items. In a classical setting the information items correspond to text documents.

With the advent of the World Wide Web, the methods of IR have been transferred to retrieval on the web. This poses different challenges and has spawned the area of Web Retrieval.

The World Wide Web (WWW) has become so common place that it is an integral part of most societies today. Millions of people around the world routinely search for information and conduct transactions using the web. Without any doubt, the WWW has been one of the fastest growing and the most pervasive phenomenon in history.

Web Mining or Web Information Retrieval (WebIR) is the process of extracting useful information from among the petabytesI of data that make up the WWW. WebIR has a history almost as long as the web itself. One of the first efforts towards WebIR was the creation of directories of web sites. An example is yahoo.com. Web directories manually organized contents of the web into a taxonomy of categories and sub categories. Web site owners who wished to be listed in the directory had to submit their site to the directory for perusal and inclusion. This rudimentary approach towards WebIR was replaced by more sophisticated mechanisms like crawler based search engines. These search engines maintained an index of words and phrases and a list of all their occurrences in the web. This index was automatically generated by programs variously called as 'spiders', 'crawlers' and 'scooters'. These agents periodically went around the web collecting and indexing web pages.

Web users can be broadly divided into three kinds based on their search strategies. These are: (a) a casual user searching the web for something that is loosely defined (b) a researcher looking for serious research level content over the web and (c) a professional looking for business intelligence by searching the web. As we see in the subsections that follow, the search strategies of each of the above user types differ substantially.

Web Information Retrieval is a pertinent topic for the present day. As the web keeps growing in size, the problem of searching the web becomes only more complex. There are a number of innovative approaches that have been proposed which hold promise. However, it remains to be seen which approach finally becomes the norm, and to what extent the web is actually used by users. The kind of WebIR technology that develops over the future would determine whether the web is destined to be a large storehouse of largely unstructured data, or is actually a huge knowledge network that offers insight to people all over the world.

WEB PERSONALIZATION ON WEB USAGE MINING

KIPSLIN CHRISTILLA. C (162MCA29)
DELLA LUKOSE (16MCA07)

Introduction

It is well known that the World Wide Web may be considered as a huge and global information centre. A web site usually contains great amounts of information distributed through hundreds of pages. Without proper guidance, a visitor often wanders aimlessly without visiting important pages, loses interest and leaves the site sooner than expected. This consideration is at the basis of the great interest about web information mining both in the academic and the industrial world.

In this section we describe our novel web usage mining strategy. It consists of two phases: in the first one a pattern analysis and classification is performed by means of an unsupervised clustering algorithm, using the registration information provided by the users. In the second one a reclassification is iteratively repeated until a suitable convergence is reached. Reclassification is used to overcome the inaccuracy of the registration information and it is accomplished by the log analysis and content management modules, based on the users' navigational behaviour. We use an unsupervised clustering procedure for partitioning the feature space built upon the user-provided data into a certain number of clusters (each one representing a class) that group together users appearing to be similar. In order to choose the optimal number of clusters, we maximize the generalization capability of the system.

In the past few years, web usage mining techniques have grown rapidly together with the explosive growth of the web, both in the research and commercial areas. In this work we present a Web mining strategy for Web personalization based on a novel pattern recognition strategy which analyses and classifies both static and dynamic features.

E-LEARNING IN WEB MINING

TEJASHWINI.S(16MCA24)
SUSHMITHA N(16MCA23)

Introduction

E-learning (also referred to as web-based education and e-teaching), a new context for education where large amounts of information describing the continuum of the teaching-learning interactions are endlessly generated and ubiquitously available. As a field of research, it is almost contemporary to e-learning. It is, though, rather difficult to define. Not because of its intrinsic complexity, but because it has most of its roots in the ever-shifting world of business. At its most detailed, it can be understood not just as a collection of data analysis methods, but as a data analysis process that encompasses anything from data understanding, preprocessing and modeling to process evaluation and Implementation. It is nevertheless usual to pay preferential attention to the Data Mining methods themselves. These commonly bridge the fields of traditional statistics, pattern recognition and machine learning to provide analytical solutions to problems in areas as diverse as biomedicine, engineering, and business, to name just a few. An aspect that perhaps makes Data Mining unique is that it pays special attention to the compatibility of the modeling techniques with new Information Technologies (IT) and database technologies, heterogeneous and complex databases. E-learning databases often fit this description. Data mining "is a process that uses statistical, mathematical, artificial intelligence and machine learning techniques to extract and identify useful information and subsequent knowledge from large databases.

1. **Data mining:** "is a process that uses statistical, mathematical, artificial intelligence and machine learning techniques to extract and identify useful information and subsequent knowledge from large databases". Data Mining can be used to extract knowledge from e-learning systems through the analysis of the information available in the form of data generated by their users. In this case, the main objective becomes finding the patterns of system usage by teachers and students and, perhaps most importantly, discovering the students' learning behavior patterns.

2. Data mining and E-learning Aims: to provide an up-to-date snapshot of the current State of research and applications of Data Mining methods in e-learning. The Cross-fertilization of both areas is still in its infancy, and even academic References are scarce on the ground, although some leading education-related Publications are already beginning to pay attention to this new field. In order to Offer a reasonable organization of the available bibliographic information According to different criteria, firstly, and from the Data Mining practitioner Point of view, references are organized according to the type of modeling Techniques used, which include: Neural Networks, Genetic Algorithms, Clustering and Visualization Methods, Fuzzy Logic, Intelligent agents, and Inductive Reasoning, amongst others. From the same point of view, the Information is organized according to the type of Data Mining problem dealt with: clustering, classification, prediction, etc. Finally, from the standpoint of the e-learning practitioner, we provide taxonomy of e-learning problems to Which Data Mining techniques have been applied, including, for instance: Students' classification based on their learning performance; detection of Irregular learning behaviors; e-learning system navigation and interaction Optimization; clustering according to similar e-learning system usage; and systems' adaptability to students' requirements and capacities.

3 Educational Data Mining (called EDM) is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in. A key area of EDM is mining computer logs of student performance
EDM include predicting student performance, and studying learning in order to recommend improvements to current educational practice. EDM can be considered one of the learning

Sciences, as well as an area of data mining.

A list of the primary applications of EDM

Analysis and visualization of data

- providing feedback for supporting instructors
- Recommendations for students
- Predicting student performance
- Student modeling
- Detecting undesirable student behaviors
- Grouping students
- Social network analysis
- Developing concept maps
- Constructing courseware
- Planning and scheduling