JYOTI NIVAS COLLEGE AUTONOMOUS

POST GRADUATE CENTER

DEPARTMENT OF MCA

E-JOURNAL

CONTENTS

USING DATA MINING TO MAKE SENSE OF CLIMATE CHANGE

AISHWARYA (17MCA01)

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."

SCREEN SCRAPER IN DATA MINING

AMALA SEELAN G V(17MCA02)

Introduction

A Screen Scraper is simply a programming script, service or other automated method that collects information from a webpage or other data source and returns it in an appropriate format for different needs. Once the data is collected, a screen scraper may include additional code for analysing the data. Screen scrapers may also be known as data scrapers, web scrapers, page scrapers, content scrapers, data harvestors, bots, agents, crawlers, spiders, indexers and other names.

Screen Scraping?

Screen Scraping is a piece of programming that mediates between legacy application programs and the modern user interfaces. Screen scraping is useful in scraping the data from SAP, MS office etc. applications used in desktop. It is designed to interact with the outdated devices and interfaces so that legacy programs can still be functional and what they contain in the form of logic and data can still be utilized. Instead of extracting/crawling data from where it is stored on the database or data files, screen scraping is important is because it gets the data from where it is displayed – the screen. It scrapes the data that was meant for the user compared to the data that is intended for another application or database.

The screen scraping application must do both of the following:

- Capture screen input and pass it on to the legacy application for processing.
- Return data from the application to the user and display it properly on the user's screen.

Screen Scraping Applications:

Here are a few areas in which screen scraping applications are used:

1. Enterprise application integration:
    - enterprise applications do not divulge the data or business rules; this integration is imperative for them. Screen scraping is preferred because it does not need to make any data structure changes and yet it is able to capture the data it needs.
    - This is particularly useful in enterprise application integration because it can make data integration between enterprise applications simple and easy.
2. Desktop analytics:
    - Desktop analytics is the process of monitoring, capturing, storing and sharing of things across applications. This is done as a part to measure and manage how processes and technology function together.
    - Screen Scraping enables to identify and work on areas of improvement in different business processes, compliance, training and usage of application. This can be accomplished by extracting, measuring, analysing and visualizing data that desktop applications generate.
3. Legacy modernization solutions:
    - legacy modernization is an inevitable and continuous process that can decrease the IT environment complexity and costs. It can also help to enhance data consistency, collaboration across platforms and increase flexibility in the process.

Conclusion:

Screen Scraping is essential for legacy applications to extend their operations. Screen scraping allows legacy applications to continue to function and remain operational.

SENTIMENT ANALYSIS IN DATA MINING

AMALI SUNITHA P (17MCA03)

Introduction:

Sentiment analysis or opinion mining is the computational study of people's opinions, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes.
The task is technically challenging and practically very useful. The Web also contains a huge amount of information in unstructured texts. This area of study is called opinion mining or sentiment analysis. It analyses people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics, and their attributes.

Tasks of Sentimental Analysis:

In this paper, we only focus on mining opinions which indicate positive or negative sentiments. The task is technically challenging and practically very useful. Opinions are important because they are key influencers of our behaviours. However, finding and monitoring opinion sites on the Web and distilling the information contained in them remains a formidable task because of the proliferation of diverse sites. The average human reader will have difficulty identifying relevant sites and accurately summarizing the information and opinions containing in them.

The basic task of opinion mining is *polarity classification*. Polarity classification occurs when a piece of text stating an opinion on a single issue is classified as one of two opposing sentiments. Reviews such as "thumbs up" versus "thumbs down," or "like" versus "dislike" are examples of polarity classification. Polarity classifications also identify pro and con expressions in online reviews and help make the product evaluations more credible.

Evolution of Opinion Mining:

Currently, opinion mining and sentiment analysis rely on vector extraction to represent the most salient and important text features. We can use this vector to classify the most relevant features. Two commonly used features are *term frequency* and *presence*. Presence is a binary-valued feature vector in which the entries indicate only whether a term occurs (value 1) or doesn't (value 0).

Multimodal Sentiment Analysis:

New sources of opinion mining and sentiment analysis abound. Webcams installed in smartphones, touchpads, or other devices let users post opinions in an audio or audio-visual format rather than in text. For a rough idea of the amount of material, consider that You-Tube users upload two days' worth of video material to its website every minute. Aside from converting spoken language to written text for analysis, the format provides an opportunity to mine opinions and sentiment.

Conclusion:

Blending scientific theories of emotion with the practical engineering goals of analysing sentiments in natural language text will lead to more bioinspired approaches to the design of intelligent opinion-mining systems capable of handling semantic knowledge, making analogies, learning new affective knowledge, and detecting, perceiving, and "feeling" emotions.

SMART SHOPPER: AN AGENT-BASED WEB-MINING APPROACH TO INTERNET SHOPPING

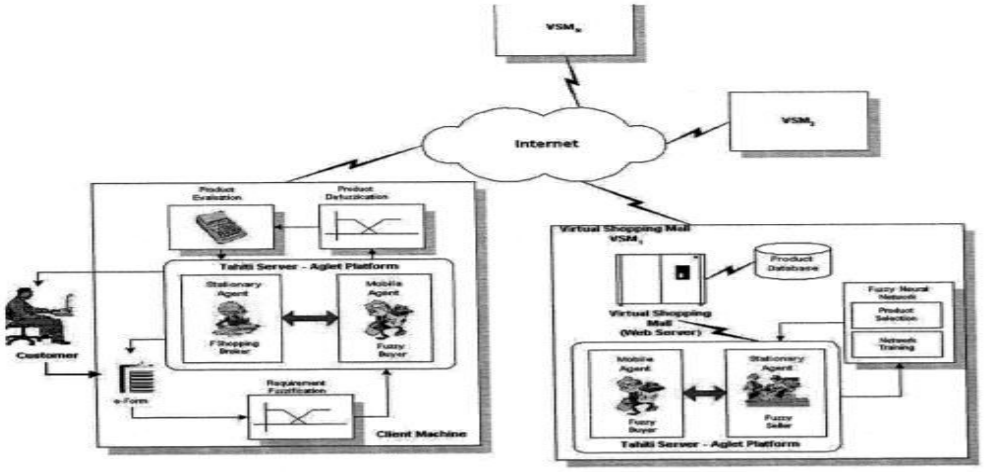
ANNAPURNA PAILA(17MCA04)

Abstract:

Technology advances in the computer industry have led to the rapid development of e-commerce applications. Nowadays, e-commerce is playing an important role in reducing costs, improving product quality, reaching new customers or suppliers, and creating new ways for selling existing products. Such a rapidly growing e-commerce environment involves communications, business transactions, services and online processing.

Understanding Smart Shopper:

Key issue in online shopping is to understand consumer behaviour. The current research suggests that such an issue should be addressed by the integration of technological, business and behavioural perspectives. There have been several models used to describe consumer buying behaviour such as the Blackwell, the Bettman, and the Howard–Sheth models. All of these models can be summarized by the following six fundamental stages of consumer buying behavior: 1) consumer requirements definition; 2) product brokering; 3) merchant brokering; 4) negotiation; 5) purchase and delivery; and 6) after-sales services and evaluation. These six stages can be implemented by using agent technologies to facilitate the web-based consumer shopping activities for e-commerce. Such agent-based systems include AuctionBot in a C2C e-auction system, BargainFinder and Jango in a B2C e-shopping system, and MAGNET in a B2B e-supply chain system



(System framework of Smart Shopper)

Methods use for it:

• FANS -The Fuzzy Buyer starts its buying activities as soon as all of the relevant information has been collected from the customer and represented in fuzzy functions. To speed up the buying process by parallelism, the Fuzzy Buyer can act as a virtual salesperson in the shopping activities.

• FPSS -The Fuzzy Seller collects all of the customer's requirements and uses a fuzzy neural network to perform product selection.

ConclusionTo overcome the limitations of the current web browsers, which lack the flexibility needed for customers to visualize products from different perspectives, we need for an autonomous,

mobile and robust system that can offer decision-making support for customers performing Internet shopping, the preliminary results from the Smart Shopper experiments etc.

CRIME DETECTION TECHNIQUES USING DATA MINING AND MACHINE LEARNING

APURVA
(17MCA05)

Introduction

Mining of data is a method of dealing with expansive data indexes to perceive outlines and set up an association to handle issues through information examination. The devices used, allow endeavours to accept future examples. Data mining is a procedure to analyse data from an informational collection to change it into a reasonable structure for additional utilization. It predicts future patterns and also enables the organization to make the learning driven decision. Generally utilized strategies for mining of data are artificial neural networks, decision tree, rule induction, nearest neighbour method and genetic algorithm. They are applied in many fields. One such interesting application is crime investigation. A crime is an unlawful activity for which a man can be penalized by law. Crime against a person is called personal crime like murder, robbery, etc. Property crime means theft of property. Crime analysis is a law implementation task which includes an organized analysis that recognizes and determines the pattern of crime. Crime can be classified into different types but, in this, we focused on four types of crime i.e. Fraud detection, traffic violence, violent crime, web crime and sexual offense. The various techniques used for different crimes have been discussed with an introduction to the concerned crime.

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 thchmiq8rse various automated data mining techniques for national security applications and local law enforcement. Particular patterns are identified 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. Techniques such as Generic algorithm, Hidden Markov Model, Naïve Bayesian, K-Mean, Neural Network, Logistic Regression, Association Rule and many more are used in crime detection. Although all of these efforts, the crime Web mining still is a highly complex task.

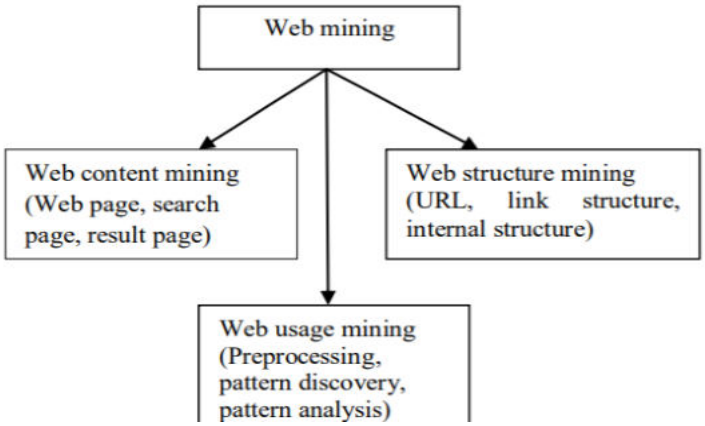
CRITICAL ANALYSIS OF SOCIAL NETWORKS WITH WEB DATA MINING

ARSHIYA ANJUM (17MCA07)

ABSTRACT: Analysis of social networks concern with the communication among users by using them as nodes of a network (graph) and their relations which are considered as network edges. Study of such type of structures place on the intersection of different fields of research: graph theory, sociology, and data mining.

INTRODUCTION: The scale of data on computers is growing at exponential rate in form of databases and files. Users need information out of these databases and files. At present, the use of Internet is increasing at rapid rate specifically associated to e-business and e-commerce applications. Data Mining is one of the kinds to support such kind of demand. Data Mining is considered as finding latent information in database. There are several challenging difficulties in data, Web and text mining research. The mining data may be structured or non-structured. Mining is of three kinds: data mining, Web mining, and text mining. Data mining concerns with structured data organized in a database while text mining concern with unstructured data and on the other side Web mining data deals the combination of unstructured and structured data.

CLASSIFICATION OF WEB MINING



• Web Content: The data actually present in the pages that conveys information to the Web users. The Web page contains multimedia data e.g. text, HTML, audio, video, images, etc. It mainly comprises: (a) Mine the data/information/ content of documents/pages, and (b) Retrieval, filtering, clustering of search results, summarization, classification/categorization, etc.

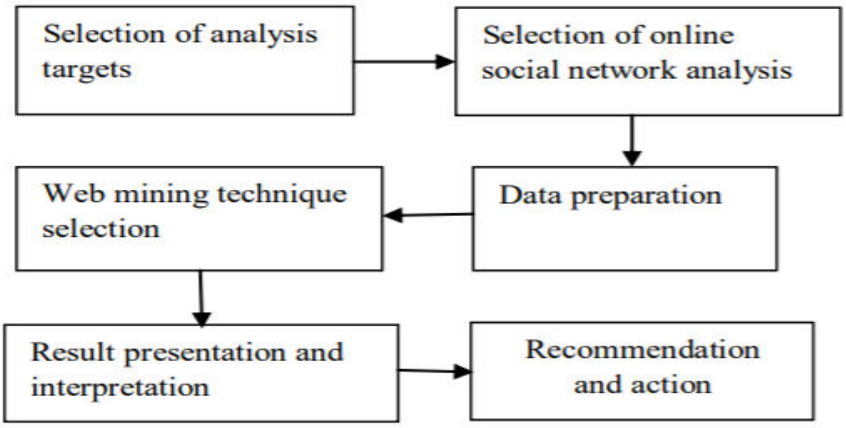
• Web Structure: The organization of the Web pages linked through hyperlinks i.e. many HTML tags used to link one page to another and one Web site to other Web site. It basically includes: (a) Study the link structure of sites and pages and sites, and (b) Authorities and hubs, detection of communities, and page ranking (Google).

• Web Usage: The data that express the usage of Web collected on proxy server, Web servers and client browser with IP address, date, time etc. It generally contains (a) Analyse surfing behavior/patterns, usage data, and (b) Site marketing and restructuring.

SOCIAL NETWORK:
A social network is generally constructed and formed by continuously daily communication among people and that is why includes different associations, such as the betweennesss, position and

closeness among groups or individuals. To understand the social structure, social relationships, social network analysis and social behaviours, it is a very useful and important technique.
For social networks analysis, the analysis objectives are mainly concentrated on Web resources, like structures, content and the user behaviours. Application of data mining techniques to the World Wide Web, called as Web mining, can be utilized for the analysis of social networks. In Web mining, key analysis objectives are from the World Wide Web, in the form of Web content mining, structure mining and usage mining.

General Web Mining Process for Social Networks Analysis steps are:



Web mining process for social networks

(i) Selection of analysis targets: The first step is the choice of the analysis targets, such as e-mail, Web, telephone communications, etc. More than one target is to be selected.
(ii) Selection of social networks analysis: The social networks analysis methodology has been choose.
(iii) Data preparation: In this step, the relevant data will be gathered for analysis and thereafter the data is to be pre-processed and cleaned to store in database.
(iv) Web mining techniques selection: Selecting the Web mining techniques or their combination to be used and then performing operation with them.
(v) Result presentation and interpretation: The analysis results after Web mining are then presented and interpreted either automatically or manually or with visualization techniques.
(vi) Recommendation and action: This is an optional step, and the process may be ended after the analysis results have been produced.

FUTURE SCOPE AND CONCLUSION:

As Web data is semi-structured /non-structured and non-homogeneous, there is a difficulty in discovery of required or unexpected knowledge information. The application of the technique- Web mining and Social networks analysis investigation carried out through the techniques of Web mining is an interesting field. However, there are many challenges to be resolve with improvement. For example, like finding communities in social networks structure, searching patterns in social networks and examining overlapping communities. We will learn to handle the challenges discussed above, besides we can also concentrate on how to utilize the Web mining techniques to some real on-line social networking Websites, such as on-line photo albums, comments and blogs.

CLUSTERING WITH EFFICIENT WEB USAGE MINING

DIVYA G M (17MCA08)

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 PERSONALIZATION USING WEB USAGE MINING

HARSHITHA M(17MCA10)

Web Personalization means making the web usage a personal experience for the user. This is done by suggesting the user some links, sites ,text ,products or messages.so the user can easily access the information he needs which will provide the user a feel that he is using his personal web. According to Web personalization can be described as any action that makes the Web experience of a user customized to the user's taste or preferences. Principal elements of Web personalization include modelling of Web objects (such as pages or products) and subjects (such as users or customers), categorization of objects and subjects, matching between and across objects and/or subjects, and determination of the set of actions to be recommended for personalization.so we can say Web personalization can be defined as any action that tailors the Web experience to a particular user, or set of users.

Web usage mining is a method in which we use the data mining techniques to identify patterns of users web usage .we make use of the clickstream data which reveals the users behavior and interest .The clickstream data means the data generated while the user do mouse clicks on various links, pictures or products. According to the process of Web personalization based on Web usage mining can consist of three stages

Data preparation and transformation

In this phase we transform raw Web log files into trans-action data which is processed with the help of Data mining tools. This phase also includes data integration from multiple sources, such as backend databases, application servers, and site content. The data obtained through this process can be divided into four

- Usage data.
- Content data.
- Structure data.
- User data.

Pattern discovery

In the pattern discovery phase we discover the user behavior patterns by applying variety of techniques on the data obtained from data preparation and transformation phase. The techniques we apply here is the clustering of data, association rules for mining and sequential pattern discovery.

Recommendation

In the recommendation phase the web personalization is done on the basis of user's active content and discovered patterns. This is done by analyzing the patterns obtained in pattern discovery phase and identifying the user's interests from them.

WEB CONTENT MINING

JOTHIS MARIA(17MCA11)

Pre-processing data before web content mining: feature selection

• Post-processing data can reduce ambiguous searching results

• Web Page Content Mining – Mines the contents of documents directly

• Search Engine Mining – Improves on the content search of other tools like search engines.

Unstructured Documents

- Bag of words, or phrase-based feature,representation Features can be boolean or frequency based

Features can be reduced using different, feature selection techniques Word stemming, combining morphological, variations into one feature

Semi-Structured Documents

-Uses richer representations for features,based on information from the document structure (typically HTML and hyperlinks)

-Uses common data mining methods (whereas⎪ unstructured might use more text mining methods)

Tries to infer the structure of a Web site or transform a Web site to become a database Better information managemen, Better querying on the Web

• Can be achieved by:

-Finding the schema of Web documents, Building a Web warehouse, Building a Web knowledge base

- Building a virtual database

• Mainly uses the Object Exchange Model (OEM) Represents semi-structured data (some structure, no

- rigid schema) by a labeled graph

• Process typically starts with manual selection of Web sites for content mining

• Main application: building a structural summary of semi-structured data (schema extraction or discovery)

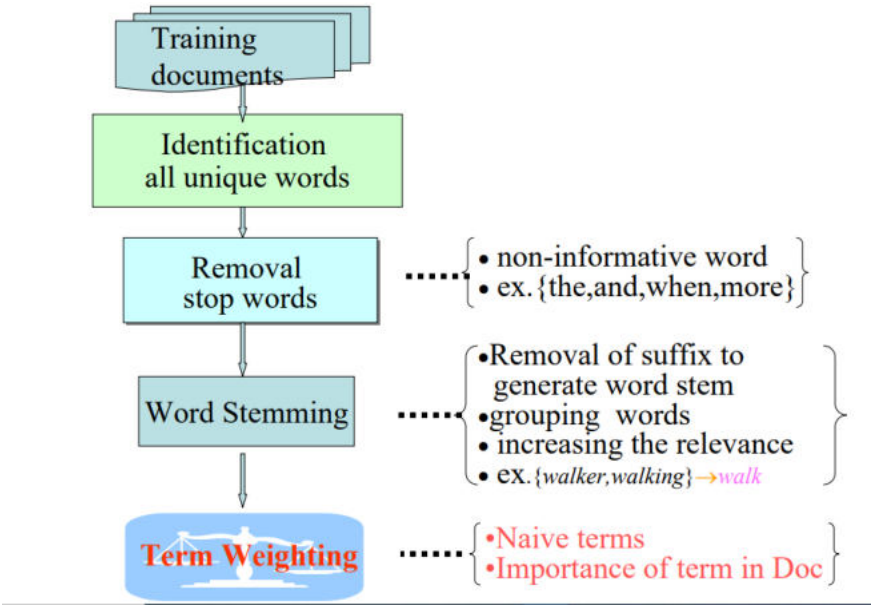
• Web content mining is related to data mining and text mining.

– It is related to data mining because many data mining techniques can be applied in Web content mining. – It is related to text mining because much of the web contents are texts. – Web data are mainly semi-structured and/or unstructured, while data mining is structured and text is unstructured.

• Applications of text mining

– Analysis of related scientific publications in journals to create an automated summary view of a particular discipline – Creation of a "relationship view" of a document collection – Qualitative analysis of documents to detect deception

Feature Extraction



Web Robots

• WWW robots (ex. spiders, wanders, crawlers, walkers, ants)

– programs that traverse WWW

– recursively retrieving pages hyperlinks by URL

– goals: automate specific Web-related tasks, e.g.

• retrieving Web pages for keyword indexing

• maintaining Web information space at local site

• Web mirroring – actually, robots never moves

WEB STRUCTURE MINING

KEERTHI K PHIRANGI(17MCA12)

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 Page rank Algorithm, HITS, Weighted Page rank 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 analysing 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.

ISSUES AND TECHNIQUES OF WEB MINING

KHUSHABOO PANDEY(17MCA13)

Abstract: The hasty growth of the web is causing the stable growth of information, leading to several problems such as an increased difficulty of extracting potentially useful knowledge. The huge amount of information available online, the World Wide Web is a fertile area for data mining research. The research in web mining aims to develop new techniques to effectively extract and mine useful knowledge or information from these web pages. Due to the heterogeneity and lack of structure of Web data, automated discovery of targeted or unexpected knowledge/information is a challenging task. In this paper, we survey the research in the area of Web mining, point out the categories of Web mining and variety of techniques used in those categories. In this paper we elicit research scope in the areas of web usage mining, web content mining, web structure mining and concluded this study with a brief discussion on data managing, querying, representation issues.

Introduction

The World Wide Web (WWW) is continuously growing with rapid increase of the information transaction volume and number of requests from Web users around the world. For web administrator' swans managers, discovering the hidden information about the users' access or usage patterns has become a necessity to improve the quality of the Web information service performances. From the business point of view, knowledge obtained from the usage or access patterns of Web users could be applied directly for marketing and management of E-business, E-services, E-searching, and E-education and so on. The following problems will be encountered during interacting with the web.

a. Finding relevant information: People either browse or use the search service when they want to find specific in-formation on the Web. When a user uses search service he or she usually inputs a simple keyword query and the query response is the list of pages ranked based on their similarity to the query. However today's a search tool have the following problems [3]. The first problem is low precision, which is due to the irrelevance of many of the search results. This results in a difficulty finding the relevant information. The second problem is low recall, which is due to the inability to index all the information available on the Web. This results in a difficulty finding the un-indexed information that is relevant.

b. Creating new knowledge out of the information available on the Web: Actually this problem could be regarded as a sub-problem of the problem above. While the problem above is usually a query-triggered process (retrieval oriented), this problem is a data-triggered process that presumes that we already have a collection of Web data and we want to ex-tract potentially useful knowledge out of it (data mining oriented). Past research [4; 5; 6] focuses on utilizing the Web as a knowledge base for decision making.

c. Personalization of the information: This problem is often associated with the type and presentation of information, since it is likely that people differ in the contents and presentations they prefer while interacting with the Web. On the other hand, the information providers could an-counter these problems, among others, when trying to achieve their goals other Web:

d. Learning about consumers or individual users: This Isa problem that specifically deals with the problem c above, which about knows what the customers do and want. Inside this problem, there are sub-problems such as mass customizing the information to the intended consumers or even to personalize it to individual user, problems related to effective Web site design and

management, problems related to marketing, etc.
Web mining techniques could be used to solve the information overload problems above directly or indirectly. However, we do not claim that Web mining techniques are the only tools to solve those problems. Other techniques and works from different research areas, such as database (DB), information retrieval (IR), natural language processing (NLP),and the Web document community, could also be used. By the direct approach we mean that the application of the Web mining techniques directly addresses the above problems. For example, a Newsgroup agent that classifies whether the news is relevant to the user. By the indirect approach we mean that the Web mining techniques are used as a part of a bigger application that addresses the above problems. For example, Web mining techniques could be used to create index terms for the Web search services.

.WEB MINING

Web mining - is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining.

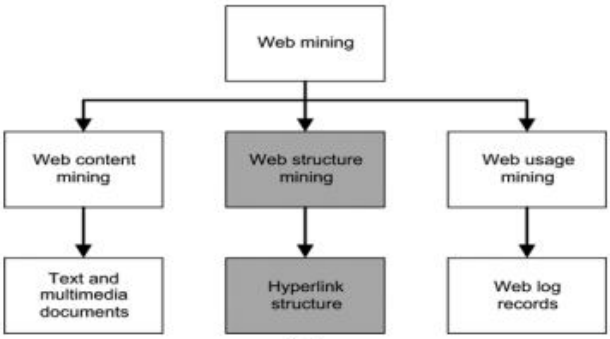


Fig-1

The web content mining mainly relates to the text and multimedia documents and web structure mining relates to the hyperlink structure and web usage mining relates to web log records.

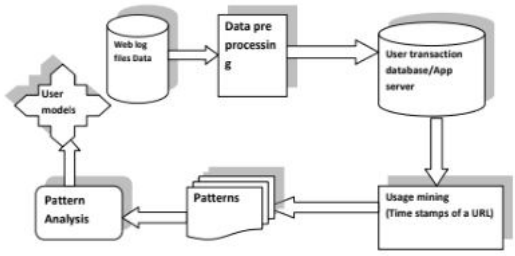


Figure-2 Web Usage Mining

The above diagram represents the overall Web usage mining process can be divided into three inter-dependent stages: data collection and pre-processing, pattern discovery, and pattern analysis. In the pre-processing stage, the clickstream data is cleaned and partitioned into a set of user transactions representing the activities of each user during different visits to the site. Other sources of knowledge such as the site con-tent or structure, as well as semantic domain knowledge from site ontology's (such as product catalogs or concept hierarchies), may also be used in preprocessing or to enhance user transaction data.

EXTRACTING AND ANALYSING WEB SOCIAL NETWORKS

KUMARI NISHA (17MCA14)

Introduction

Recently, online social networks have gained significant popularity and are now among the most popular sites on the Web. A social network is a structured representation of the social actors (nodes) and their interconnections (ties) and form social groups that share common interests. Online social networks have emerged as a powerful tool for personal communication and interaction. Web based communities have become important places for people to seek and share knowledge and expertise. Online social networks are organized around users in contrast to the Web which is largely organized around content.

Social network extraction

The World Wide Web (WWW) has become a popular medium to distribute information today. Data on the Web is rapidly increasing and is huge, diverse and dynamic so information users could encounter problems like: finding relevant information; creating new knowledge out of the information available on the web; personalization of information; learning about consumers or individual users, while interacting with the Web. Exponential growth of public information on the Web in a set of interlinked heterogeneous sources has made the information search a challenging task. Search engines are the most widely used tools for searching information on the Web, but the general approaches to analyse this information cannot integrate different sources.

Web Mining Techniques for Social Network Extraction

Web mining deals with the discovery and extraction of useful information from the Web. Web mining can be classified into Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM). WCM analyses the contents on the Web, WSM deals with the links and structure of websites, and WUM can be used to analyses how websites have been used.

Web mining is very useful in online social network analysis and extraction. WCM can be used for a number of purposes such as categorization or classification of documents on an online social networking website, analysing users' reading interests, determining their favourite content, etc.

WUM provides usage data and user communications logs on an online social networking website. This data can be transformed into relational data for social-networks construction. WSM is very useful for extracting online social networks by extracting the links from WWW, e-mail or other sources.

In addition, it can also be used to analyses path length, reachability or to find structural holes. For most online social networks analyses, the three types of Web mining can't work alone and it is usually necessary to utilize all three types of Web mining techniques together.

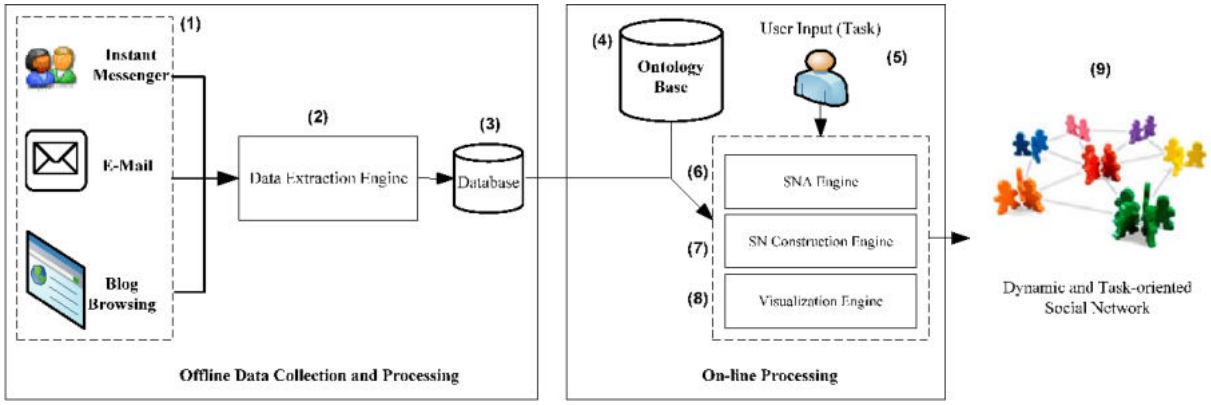
Web Information Sources

The relationship information for extracting of social networks is obtained from various online sources. The sources that have been used in these studies are Web pages, e-mail communications,

instant messaging, Internet relay chats, blogs, online social networking sites, news, web albums, etc. The different social network extraction techniques can be classified into following six categories on the basis of information source they use.

- Web based Social Network Extraction.
- E-mail Communication based Social Network Extraction.
- Instant Messaging/ Chat based Social Network Extraction.
- Blogs based Social Network Extraction.
- Online Social Networking Sites Data based Social Network Extraction.
- Multi-Source Data based Social Network Extraction.

Social network extraction methods



Conclusion

We classified and discussed various automatic methods for social network extraction. These methods have an edge over the manual methods where extracting attributes involved a lot of interaction with the profile owner, e.g., questionnaires and interviews. We also proposed a general framework that can be used for building of social network extraction systems. But several challenges like data sampling, combination of different types of web mining techniques, information representation on the Web are still to be addressed properly. It is necessary to design efficient tools and techniques for data extraction from the Web because it becomes difficult for end users to find useful data because of a number of issues. Information representation is one of them.

A STUDY ON WEB MINING TECHNIQUES IN SOCIAL MEDIA

MARY HARSHITHA A(17MCA15)

Abstract

Data mining is the procedure of examining pre-existing databases to generate new useful information by using some strategies and implementing particular operations. Accessing social network sites such as Twitter, Facebook, LinkedIn and Google+ through the internet has become very useful and daily part of our life. People are becoming more interested in and depending on social network for information, news and opinion of other users on diverse subject matters. Social Media mining is the process of representing, analysing and fetching actionable patterns from raw data in social media by using many techniques to conquer the problems in social media.

Introduction

Data mining is the collection of techniques for efficient discovery of previously unknown, valid, novel, useful and understandable patterns in large databases. It is usually used by business intelligence organizations, and financial analysts, but it is increasingly used in the science to get information from the enormous data sets generated by modern experimental and observational methods.

Social Data Mining

Social data mining introduces basic concepts and principle algorithm suitable for exploring enormous social media data, it discusses theories and methodologies from different direction such as computer science, machine learning, social network analysis, data mining, optimization and mathematics. Social media asserts to the growth of these social networks in which individual collaborate with one another through friendship, emails, BlogSpot and many other mechanisms. Social media mining intents to make sense of these individuals enclosed in networks.

Various Techniques of Data mining in Social Media

There are many different types of techniques have been developed to overcome the problems such as size, noise, and dynamic nature of social media data. Due to different types of data and massive volume of data in the social media, it requires an automatic data processing in order to analyse it within a given time span. Different types of data mining techniques are as follows.

1. Unsupervised classification

We can easily decide a review as ‘thumbs-up’ or ‘thumbs down’ by using unsupervised learning. This type of marking can be done locating the phrases including an adjective or adverb. We can estimate the semantic orientation of every phrase by using PMI-IR followed by the grouping of the review by using the mean semantic orientation of the phrase.

1.1 Sentiment lexicon

Sentiment lexicon is a collection of sentimental words that are used by reviewers in their expressions. Sentiment lexicon is a catalogue of the common words that intensify data mining techniques. Different aggregation of sentiment lexicon can be created for assortment of subject matters. For example, sentimental words used in politics are often different those used in sports.

Sentiment orientation

Sentiment orientation can be positive, negative, or neutral (no opinion). It might be immense for the future buyers to make the decision regarding the purchase of a product by tracking usable reviews which are attracted by the widespread products. Semantic orientation is also used by the application developers for their application ranking so that they could see the reviews presented by the users.

Opinion definition and summarization

These are the important techniques granting opening. Opinion definition can be discovered in a text, sentence or the document's topic, and it can also occupy the whole document. Opinion extraction is difficult for summarization and tracking of any document. Using this technique, the biased (fixed views) part is explored in the texts, and documents. It is required to aggregate the opinion since all the opinions fetched in the document are not as a direct result of consequence concerning the topic under analysis.

Basic clustering technique

Clustering can be considered the most important unsupervised learning problem; it deals with finding a structure in a collection of unlabelled data. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. Clustering techniques can be applied in many fields, for instance: Marketing, Biology, Libraries, Insurance, City-planning, Earthquake studies, and WWW (World Wide Web). Clustering techniques involves four most used clustering algorithms; K-means, Fuzzy C-means, Hierarchical clustering, Mixture of Gaussians. So that, Kmeans is an exclusive clustering algorithm, Fuzzy C-means is an overlapping clustering algorithm, Hierarchical clustering is obvious and lastly, Mixture of Gaussian is a probabilistic clustering algorithm

1.2 Opinion extraction

This technique is compulsory in order to aim that chunk of the document including genuine opinion. An individual's opinion regarding a skilled subject does not matter unless that particular individual has mastered that specific domain. However, the use of both opinion extraction and summarization is essential because of the opinion from many people. The massive number of people giving their opinion regarding a certain subject, it will be more significant to take out that particular.

.

INTRODUCTION TO INTEGRATING WEB MINING WITH NEURAL NETWORK

NANDINI D R (17MCA16)

ABSTRACT

The World Wide Web is huge, unstructured, universal and heterogeneous. In recent years the growth of the World Wide Web exceeded all expectations. Today there are several billions of HTML documents, pictures and other multimedia files available via internet and the number is still rising. But considering the impressive variety of the web, retrieving interesting content has become a very difficult task. Web usage mining is one of the technique of web mining is very useful to discover knowledge from secondary data obtained from the interaction from users with the web. The web usage mining is very essential for effective website. In this paper we give basic idea about web mining, neural network and GNG algorithm.

INTRODUCTION

The Application of data mining techniques to the World Wide Web referred as web mining. We make use of the web in several ways. For example, finding relevant information, discovering new knowledge from the web, personalized web page synthesis, learning about individual users etc. Web mining techniques provides a set of techniques which provide solutions to different problems. However, web mining techniques are not the only tools to handle these problems. Other related techniques from different research areas such as database (DB), information retrieval (IR) and natural language processing (NLP) can also be used. When we see web mining in terms of data mining it have three interest of operations say clustering (e.g. finding natural groupings of users, pages, etc.), associations (e.g. which URLs tend to be requested together) and sequential analysis (e.g. the order in which URLs tend to be accessed). As in most real world problems the clusters and associations in web mining do not have clear cut boundaries and often overlap considerably. Web mining techniques can be categorized as web content mining, web structure mining and web usage mining.

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

- Web content mining: It studies the search and retrieval of information on the web. Web content mining future can be divided as web page content mining and search result mining. It has to do with the retrieval of content available on the web into more structure forms as well as its indexing for easy tracking information locations.
- Web structure mining: It focuses on the structure of the hyperlinks (inter document structure) within the web. 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. Then it focuses on the identification of authorities.
- **Web usage mining**: is the process of identifying browsing patterns by analysing the user's navigational behaviour. This information takes as input the usage data i.e. the data residing in the web server logs, recording the visits of the users to a web site.

CONCLUSION:

Web usage mining is used in many areas such as e-Business, e-CRM, e-Services, e-Education, e-Newspapers, e-Government, advertising, Digital Libraries, marketing, bioinformatics and so on. One of the open issues in data mining, in general and Web Mining, in particular, is the creation of intelligent tools that can assist in the interpretation of mined knowledge. Clearly, these tools need to have specific knowledge about the particular problem domain to do any more than filtering based on statistical attributes of the discovered rules or patterns.

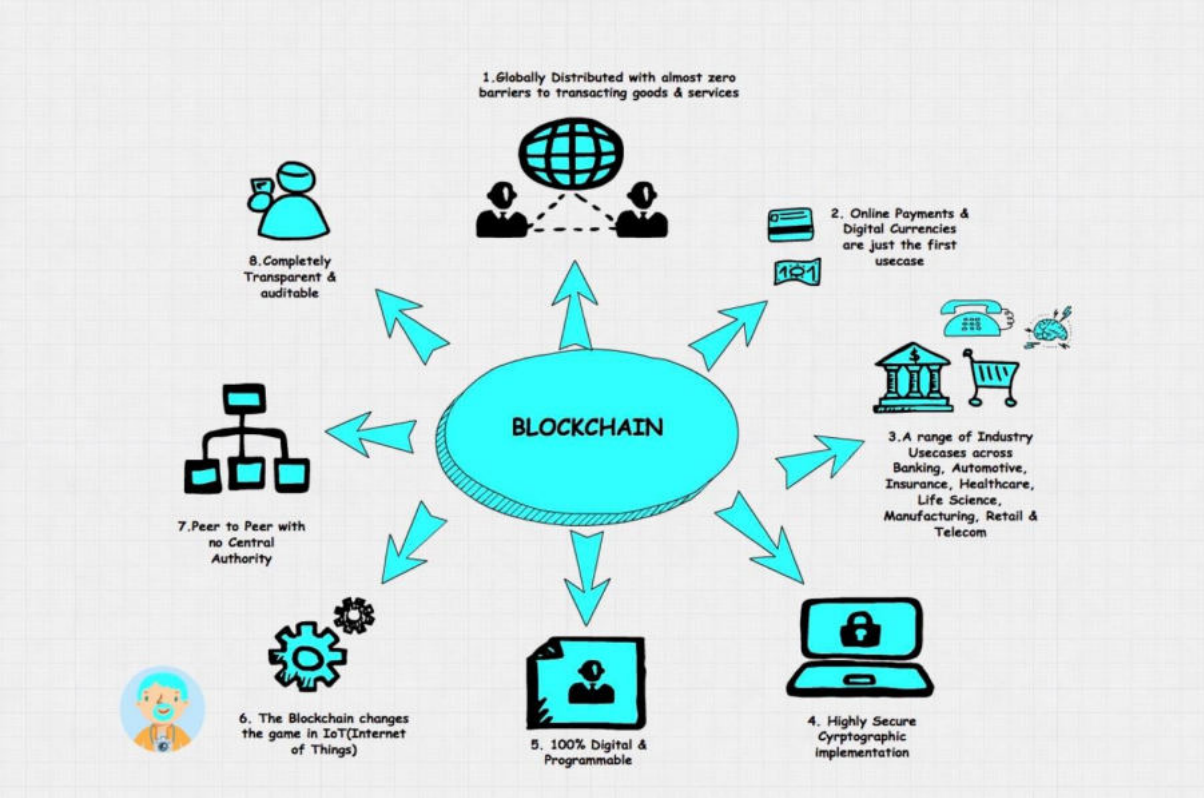
BLOCK CHAIN IN WEB DATA MINING

POOJASHREE N(17MCA18)

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.

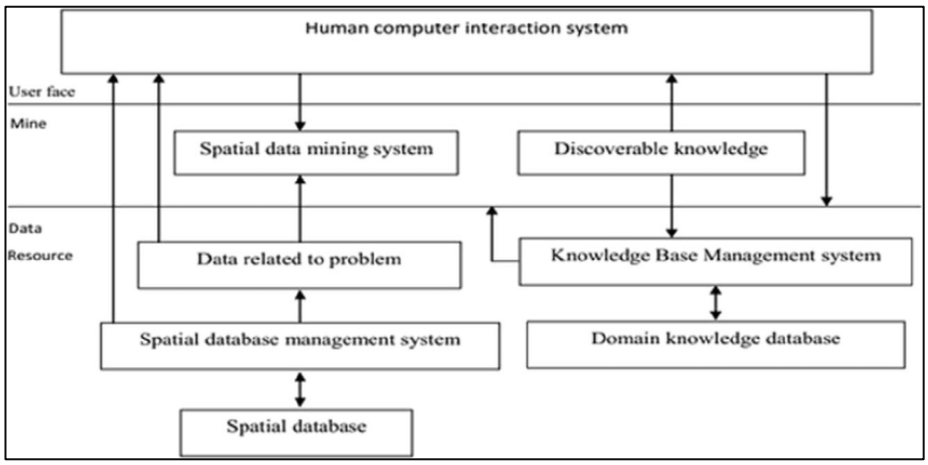
As we have discussed about few promising block chain projects that will help transition our web from 2.0 to 3.0. It will be possible that these projects will take some time before we see some good results and these problem can be solved when the developers will completely understand the block chain technology.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

SPATIAL DATA MINING

R BHAVYASHREE(17MCA21)

Spatial data mining it's the mining knowledge from large amounts of spatial data, is a highly demanding field because huge amounts of spatial data have been collected in various applications, ranging from remote sensing, to geographical information systems (GIS), computer cartography, environmental assessment and planning, etc. The collected data far exceeded human's ability to analyse. Recent studies on data mining have extended the scope of data mining from relational and transactional databases to spatial databases. This paper summarizes recent works on spatial data mining, from spatial data generalization, to spatial data clustering, mining spatial association rules, etc.

Spatial Data Mining Structure



The spatial data mining can be used to understand spatial data, discover the relation between space and the non-space data, set up the spatial knowledge base, excel the query, reorganize spatial database and obtain concise total characteristic etc. The system structure of the spatial data mining can be divided into three layer structures mostly. The customer interface layer is mainly used for input and output, the miner layer is mainly used to manage data, select algorithm and storage the mined knowledge, the data source layer, which mainly includes the spatial database and other related data and knowledge bases, is original data of the spatial data mining.

OBJECT ORIENTED DATABASE

Object oriented databases are based on the object oriented programming paradigm; each entity is considered as an object. Data and code relating to an object are encapsulated into a single unit. Each object has associated with it the following:

a. Set of variables that describes the object.

b. Set of messages that the object can use to communicate with other objects, or with the rest of the database system.

c. Set of methods, where each method holds the code to implement a message.

E-LEARNING IN WEB MINING

RAKSHTHA G L (17MCA22)

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, pre-processing and modelling 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 modelling 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 behaviour 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 modelling 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 behaviours; 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 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.

INFORMATION AND PATTERN DISCOVERY ON WORLD WIDE WEB

ROJA RANI K (17MCA23)

With the explosive growth of information sources available on the World Wide Web, it has become increasingly necessary for users to utilize automated tools to find the desired information resources, and to track and analyze their usage patterns. These factors give rise to the necessity of creating server-side and client-side intelligent systems that can be effectively mined for knowledge. Web mining can be broadly defined as the discovery and analysis of useful information from the World Wide Web. This describes the automatic search of information resources available on-line, i.e. Web content mining, and the discovery of user access patterns from Web servers, i.e., Web usage mining.

Web Content Mining

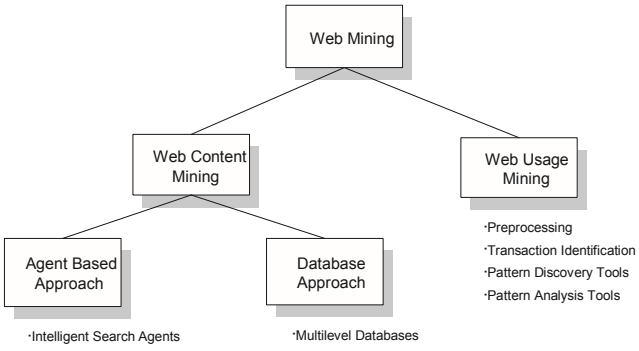
The lack of structure that permeates the information sources on the World Wide Web makes automated discovery of Web-based information difficult. Traditional search engines do not provide structural information nor categorize, filter or interpret documents.

These factors have prompted researchers to develop more intelligent tools for information retrieval, such as intelligent Web agents, and to extend data mining techniques to provide a higher level of organization for semi-structured data available on the Web.

Web Usage Mining

Web usage mining is the automatic discovery of user access patterns from Web servers. Organizations collect large volumes of data in their daily operations, generated automatically by Web servers and collected in server access logs.

Analysing such data can help organizations determine the life time value of customers, cross marketing strategies across products, and effectiveness of promotional campaigns, among other things. It can also provide information on how to restructure a Web site to create a more effective organizational presence, and shed light on more effective management of workgroup communication and organizational infrastructure.



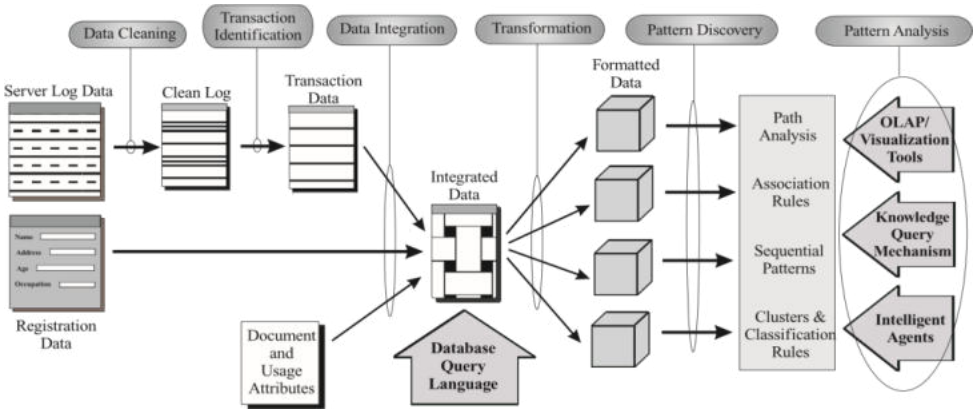
Pattern Discovery from Web Transactions

Analysis of how users are accessing a site is critical for determining effective marketing strategies and optimizing the logical structure of the Web site. Because of many unique characteristics of the client-server model in the World Wide Web, including differences between the physical topology of Web repositories and user access paths, and the difficulty in identification of unique users as well as user sessions or transactions, it is necessary to develop a new framework to enable the mining process. Specifically, there are a number of issues in pre-processing data for mining that must be addressed before the mining algorithms can be run. These include developing a model of access log

data, developing techniques to clean/filter the raw data to eliminate outliers and/or irrelevant items, grouping individual page accesses into semantic units (i.e. transactions), integration of various data sources such as user registration information, and specializing generic data mining algorithms to take advantage of the specific nature of access log data.

Analysis of Discovered Patterns

The discovery of Web usage patterns, carried out by techniques described earlier, would not be very useful unless there were mechanisms and tools to help an analyst better understand them. Hence, in addition to developing techniques for mining usage patterns from Web logs, there is a need to develop techniques and tools for enabling the analysis of discovered patterns. These techniques are expected to draw from a number of fields including statistics, graphics and visualization, usability analysis, and database querying.



Conclusion

Proposed a definition of Web mining, and developed a taxonomy of the various ongoing efforts related to it. Also provided a general architecture of a system to do Web usage mining, and identified the issues and problems in this area that require further research and development.

Neuro-Fuzzy Based Hybrid Model for Web Usage Mining

SANJANA K R (17MCA24)

Web Usage mining consists of three main steps: Pre-processing, Knowledge Discovery and Pattern Analysis. The information gained from the analysis can then be used by the website administrators for efficient administration and personalization of their websites and thus the specific needs of specific communities of users can be fulfilled and profit can be increased. Also, Web Usage Mining uncovers the hidden patterns underlying the Web Log Data. These patterns represent user browsing behaviours which can be employed in detecting deviations in user browsing behaviour in web based banking and other applications where data privacy and security is of utmost importance. A neuro-fuzzy based hybrid model is employed for Knowledge Discovery from web logs.

In Web Usage Mining browsing patterns of users are analysed to extract useful information. Business communities make use of the discovered knowledge to increase the profit by personalizing the web sites for the customer thereby improved customer satisfaction. Web Usage Mining discovers interesting usage patterns hidden in web data. Usually, Web Usage Mining consists of Pre-Processing, Knowledge Discovery and Pattern Analysis. Web Server Logs serves as the input to the Web Usage Mining process. The Web Log Data is unstructured and noisy and ambiguous. a hybrid model based on neuro – fuzzy clustering is implemented to efficiently cluster the users of website based on similar browsing patterns

Knowledge discovery using neuro – fuzzy clustering algorithm

Once the user sessions have been identified, a clustering process is applied in order to group similar sessions in the same cluster. Each cluster includes users exhibiting a common browsing behaviour and hence similar interests. In the earlier work, the Fuzzy C-Means Clustering algorithm was employed to cluster the user sessions. Artificial neural network seeks to emulate the architecture and information representation patterns of the human brain. Artificial neural networks are designed as per the target to be accomplished. Patterns are presented at the input, which are associated with the output nodes with differential weights. An iterative process is followed to adjust the weights between the input nodes and the output nodes until a termination criterion is satisfied. This process of weight adjustment, called learning, provide, continuous learning or artificial learning capability to the system, which can be either supervised or unsupervised learning, provide, continuous learning or artificial learning capability to the system, which can be either supervised or unsupervised learning in artificial neural networks. The supervised learning demands an output class declaration for each of the inputs. In hybrid model based on neural networks and fuzzy clustering to cluster users according to the browsing patterns. First, a sample set of pre-processed web log data is clustered using Fuzzy C Means Clustering algorithm. Then, the input of clustering algorithm is given as input to neural network and the output of clustering algorithm is given as target output and the training of neural network is done. Mean Square Error – the average squared error between the network outputs and the target outputs is used as the performance measure

WORLD TOWARDS ADVANCE WEB MINING

SWEETY LENKA(17MCA26)

Abstract

With the advent of the World Wide Web and the emergence of e-commerce applications and social networks, organizations across the Web generate a large amount of data day-by-day. The abundant unstructured or semi-structured information on the Web leads a great challenge for both the users, who are seeking for effectively valuable information and for the business people, who needs to provide personalized service to the individual consumers, buried in the billions of web pages. To overcome these problems, data mining techniques must be applied on the Web.

Introduction

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. It tries to discovery the useful information from the secondary data derived from the interactions of the users while surfing on the Web. Usage data captures the identity or origin of Web users along with their browsing behaviour at a web site. It deals with studying the data generated by web surfer's sessions or behaviours. Since the web content and structure mining utilize the real or primary data on the web. On the contrary, web usage mining mines the secondary data derived from the interactions of the users with the web. The secondary data includes the data from the proxy server logs, browser logs, web server access logs, user profiles, user sessions, user queries, registration data, bookmark data, mouse clicks and scrolls, cookies and any other data which are the results of these interactions.

Web Mining Application Areas

Web mining is an important tool to gather knowledge of the behaviour of Websites visitors and thereby to allow for appropriate adjustments and decisions with respect to Websites actual users and traffic patterns. Along with a description of the processes involved in Web mining states that Website Design, Web Traffic Handling, e-Business and Web Personalization are four major application areas for Web mining. These are briefly described in the following sections.

1. Website Design

The content and structure of the Website is important to the user experience/impression of the site and the site 's usability. The problem is that different types of users have different preferences, background, knowledge etc. making it difficult (if not impossible) to find a design that is optimal for all users. Web usage mining can then be used to detect which types of users are accessing the website, and their behaviour, knowledge which can then be used to manually design/re-design the website, or to automatically change the structure and content based on the profile of the user visiting it.

2. Web Traffic Handling

The performance and service of Websites can be improved using knowledge of the Web traffic in order to predict the navigation path of the current user. This may be used for cashing, load balancing or data distribution to improve the performance. The path prediction can also be used to detect fraud, break-ins, intrusion etc.

3. e-Business

E-commerce or e-commerce, is trading in products or services using computer networks, such as the Internet. Electronic commerce draws on technologies such as mobile commerce, electronic funds transfer, supply chain management, Internet marketing, online transaction processing, electronic data interchange (EDI), inventory management systems, and automated data collection systems. Modern electronic commerce typically uses the World Wide Web for at least one part of the transaction's life cycle, although it may also use other technologies such as e-mail. E-commerce businesses may employ some or all of the following:

- Online shopping web sites for retail sales direct to consumers
- Business-to-business buying and selling
- Gathering and using demographic data through web contacts and social media
- Business-to-business electronic data interchange
- Engaging in pretail for launching new products and services.

4. Web Personalization

Personalizing web means, for a given query, the web search engine produces different ***SERPs*** or reorganizes the SERPs differently for different users. For this, the intuition of the user is captured by the usage patterns. The ***search engine results page*** (SERP) is the actual result returned by a search engine in response to a keyword query. The SERP consists of a list of links to web pages with associated text snippets. The SERP rank of a web page refers to the placement of the corresponding link on the SERP, where higher placement means higher SERP rank. Web Personalization Process contains:

1. Define an Audience based on the Visitor's

2. Deliver Personalized Content

3. Optimize and Test to Perfection

Conclusion

Web mining is a rapid growing research area. As the Web has become a major source of information, techniques and methodologies to extract quality information is of paramount importance for many Web applications and users. Web mining and knowledge discovery play key roles in many of today's prominent Web applications such as e-commerce and computer security.

INFORMATION FITERING USING WEBMINING

TEJASWINI G (17MCA27)

ABSTRACT:
Information filtering is the method to identify the most important results from a list of discovered frequent set of data items for which you can make use of web mining.

INTRODUCTION:
Information filtering deals with the delivery of information that the user is likely to find interesting or useful. An information filtering system assists users by filtering the data source and deliver relevant information to the users. When the delivered information comes in the form of suggestions an information filtering system is called a recommender system. Because users have different interests the information filtering system must be personalized to accommodate the individual user's interests. This requires the gathering of feedback from the user in order to make a user profile of his preferences.

TYPES OF FILTERING:

- **Content-based filtering**: Also referred to as cognitive filtering, recommends items based on a comparison between the content of the items and a user profile. The content of each item is represented as a set of descriptors or terms, typically the words that occur in a document. The user profile is represented with the same terms and built up by analysing the content of items which have been seen by the user..
  Relevance feedback, genetic algorithms, neural networks, and the Bayesian classifier are among the learning techniques for learning a user profile.

- **Collaborative filtering**: Also referred to as social filtering, filters information by using the recommendations of other people. It is based on the idea that people who agreed in their evaluation of certain items in the past are likely to agree again in the future. A person who wants to see a movie for example, might ask for recommendations from friends. The recommendations of some friends who have similar interests are trusted more than recommendations from others. This information is used in the decision on which movie to see.
- **Hybrid system** : A hybrid approach combines the two types of information while it is also possible to use the recommendations of the two filtering techniques independently.

CONCLUSION:
Alternative approaches for filtering information have been proposed as well. Demographic filtering systems for example use demographic information such as age, gender and education to identify the types of users that like a certain item. Economic filtering systems select items based on the costs and benefits of producing and viewing them. An example of economic filtering are systems that adaptively schedule banner advertisements on the internet. Ad systems exists that learn to display ads that will yield the highest possible click-through rate based on the past behavior of the user. By directing ads to a more targeted population it could help internet providers and advertising agents increase their ad revenues.

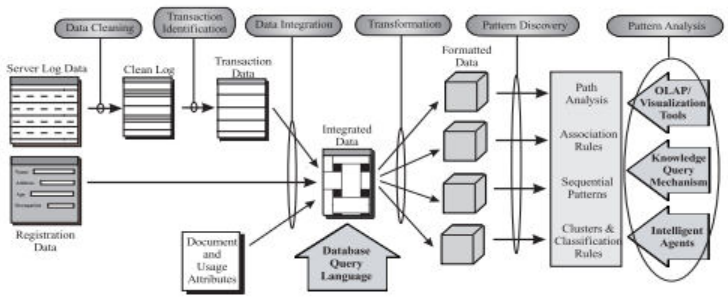
Web Mining: Information and Pattern Discovery on the World Wide Web

Varsha Sahay(17MCA28)

Abstract:
Application of data mining techniques to the World Wide Web, referred to as Web mining, has been the focus of several recent research projects and papers. However, there is no established vocabulary, leading to confusion when comparing research efforts. The term Web mining has been used in two distinct ways. The first, called Web content mining in this paper, is the process of information discovery from sources across the World Wide Web. The second, called Web usage mining, is the process of mining for user browsing and access patterns

Understanding Pattern Discovery from Web Transactions: -

As discussed analysis of how users are accessing a site is critical for determining effective marketing strategies and optimizing the logical structure of the Web site. Because of many unique characteristics of the client-server model in the World Wide Web, including differences between the physical topology of Web repositories and user access paths, and the difficulty in identification of unique users as well as user sessions or transactions, it is necessary to develop a new framework to enable the mining process. Specifically, there are a number of issues in pre-processing data for mining that must be addressed before the mining algorithms can be run. These include developing a model of access log data, developing techniques to clean/filter the raw data to eliminate outliers and or irrelevant items, grouping individual page accesses into semantic units (i.e. transactions), integration of various data sources such as user registration information, and specializing generic data mining algorithms to take advantage of the specific nature of access log data.



(A General Architecture for Web Usage Mining)

Methods use for it:-

• Pattern Analysis Tools. Once access patterns have been discovered, analysts need the appropriate tools and techniques to understand, visualize, and interpret these patterns, e.g. the WebViz system.

• Pattern Discovery Tools. The emerging tools for user pattern discovery use sophisticated techniques from AI, data mining, psychology, and information theory, to mine for knowledge from collected data. For example, the WEBMINER system.

Conclusion:-

The term Web mining has been used to refer to techniques that encompass a broad range of issues. However, while meaningful and attractive, this very broadness has caused Web mining to meandifferent things to different people and there is a need to develop a common vocabulary.

FRAUD DETECTION IN DATAMINING

Y HARI CHANDANA(17MCA29)

Abstract: Fraud detection is a technique of identifying prohibited acts that are occurring around the world. It defines the skilled impostor, formalizes the key forms and sub forms of recognized frauds and reveals the gathered data nature. To analyze fraud patterns from data this paper represent preferred data mining techniques. Now a days Data mining is widely used concept, that can be used everywhere for analysing data patterns.

INTRODUCTION: Data mining is about discovering new patterns which are unknown before, statistically reliable and process able from data. Data mining is a field which is concerned to understanding data patterns from huge datasets. We can say that the aim is to find out new patterns in data.

II. TECHNIQUES USED FOR ANALYSE AND DETECT FRAUD PATTERNS:

In Data mining generally four modules of task exist:

1) **Classification -** Data are arranged into predefined groups with the use of different algorithms. Classification is the grouping of data in predefined classes. We can also say supervised classification, the classification which uses given class to arrange the objects in the data group. Classification techniques generally use a training set for objects those are previously grouped by known class. Algorithms of classification learn from the training set and create a structure for this. This structure is used for groping new objects. Different classification techniques used for fraud data patterns according to their nature.

2) **Clustering –** It is also like classification but predefined classes do not exist there, so that clustering algorithms attempt to do similar objects together. Like classification technique, clustering is the association of data in classes, but unlike classification, in clustering, classes are not predefined. We can say that Clustering is an unsupervised classification, because the classification is not based on previously known classes. Clustering approaches based on the principle of similarity maximization among intra-class objects and similarity minimization among inter-class objects.

3) **Regression –** In this process we try to obtain a function which models the data of the minimum error. A general term is to use Genetic Programming.

4) **Association rule –** It is used to find relationship among data objects. This analysis of association of objects is the discovery commonly called association rules. It observes the frequency sets occurring simultaneously in transactional database. It is based on two threshold values support and confidence. Support, identifies the frequent item sets and confidence is the conditional probability that an item appears in a transaction when another item appears. These are the basic terms used in the data analysis using data mining but now we understand techniques used according to fraud patterns.

CONCLUSION:

We studied different fraud detection data mining techniques according to different areas. Data mining is a well-known zone of analysing, predicting and defining rules from the large amount of data and finding true, previously unknown patterns. This mainly focuses on data mining techniques as impressive approach for fraud patterns detection in every area.

MINING E-GOVERNANCE USING DATA WAREHOUSING

AFNAAN K (17MCA30)

Introduction

The basic requirements of good governance are derived from the fact that the laws and methods are well defined, transparent and easily understandable by people. To provide such good governance in a developing country like India is a challenge itself as most of the people are not educated or are not economically strong. The challenge becomes larger in developing countries as the democratic methods are used in forming the governments. In number of cases the rules and procedures defined in the constitution themselves become obstacle in the path of governance due to absence of transparency and procedural clarities. The solution to the above problems lies in developing a mechanism that is interactive, fast and provides a clear repository of guidelines that can be used in effective decision making for both the government and the people. E-Governance is the mode that has large number of advantages in implementing easy, transparent, fair and interactive solutions within minimum time frame.

E-GOVERNANCE

E-Governance involves collection of technology based processes that involves greater interaction between government and citizens and hereby have highly improved delivery of public services . E-Governance is based on the effective utilization of information and communication technologies (ICT) with major objectives of making public representatives more transparent, accountable and effective by providing improved information and service delivery and enhanced participation of people in day to day activities

MINING E-GOVERNANCE DATA WAREHOUSE

Data warehouse is used for collecting, storing and analysing the data to assist the decision making process. Data mining can be applied to any kind of information repository like data warehouses, different types of database systems, World Wide Web, flat files etc.. Therefore, data warehousing and data mining are best suited for number of applications based on e-Governance in G2B (Government to Business), G2C (Government to Citizen) and G2G (Government to Government) environment. In order to have effective implementation there should be solid Data Warehouse on data collected from heterogeneous reliable sources The subcategories of e-government are described in Table 1.

Parties of communication	Dominant Characteristics	Definition	Example
Government-to-Government (G2G)	Communication, coordination, standardization of information and services	E-administration	Establishing and using a common data warehouse
Government-to-Citizen (G2C)	Communication, transparency, accountability, effectiveness, efficiency, standardization of information and services, productivity	E-government	Government organization Web Sites, e-mail communication between citizens and government officials
Government-to-Business (G2B)	Communication, collaboration, commerce	E-government, E-Commerce, E-collaboration	Posting government bids on the Web, e-procurement, e-partnerships
Government-to-Civil Society Organization (G2SC)	Communication, coordination, transparency, accountability	E-governance	Electronic communication and coordination efforts after a disaster
Citizen-to-Citizen (C2C)	Communication, coordination, transparency, accountability, grassroots organization	E-governance	Electronic discussion groups on civic issues

a) Phase I: The e-governance is made available online i.e. providing relevant information to the people (G2C & G2B). Earlier government websites were quite similar to brochure or leaflet but there is paradigm shifts as more and more information is made available on web. The major advantage is that government information is publicly accessible; processes are described and become more transparent, which improves democracy and service.

b) Phase II: It involves the communication between policy makers and the public (G2C & G2B). Public can get their queries solved via e-mail, use search engines, and download forms and documents. The applications can be processed online at very fast rate. Internally lans, intranets and e-mail are used to communicate and exchange data by various government departments (G2G).

c) Phase III: The complexity of transactions increases in third phase. Complete transactions can be performed at the leisure of house. Extending/renewal of licenses, application for visa and passports filing property tax, filing income tax, and online voting are common examples. This phase handles complex queries by use of security and personalization issues. E.g. Digital signatures will be mandatory to have legal transfer of services. The government has also made e-procurement compulsory for all procurements above Rs 5,000/-

d) Phase IV: The major goal is to provide single counter by integrating all information systems. The employees in various government departments have to work in the coordinated manner to have cost savings, efficiency and most importantly highest customer satisfaction.

A WEB MINING APPROACH TO HYPERLINK SELECTION FOR WEB PORTALS

HARSHITHA G (172MCA31)

As the size and complexity of websites expands dramatically, it has become increasingly challenging to design websites on which web surfers can easily find the information they seek. To address this challenge, we introduce a new problem in this area, hyperlink selection, and present a web mining based approach, Link Selector, as a solution.

The homepage of a website is one type of portal page. Homepages which guide users to locate the information they seek easily create a good first impression and attract more users, while homepages which make information searching difficult result in a bad first impression and corresponding user loss [Nielson and Wagner 1996]. A default web portal is another type of portal page. Recently, web portals that serve as a personal entrances to websites have attracted more and more attention. Universities such as UCLA have built educational web portals corporations such as Yahoo! have developed commercial web portals . For practical purposes, portal service providers (e.g., Yahoo!) provide portal users with a standard default web portal, which the users can personalize (e.g., add or remove hyperlinks from the default web portal). As the first version of a web portal encountered by portal users, the default web portal plays an important role in the success of a web portal. Moreover, according to My Yahoo!, most users never customize their default web portals [Manber et al. 2000]. This finding makes the default web portal even more critical. A portal page consists of hyperlinks selected from a hyperlink pool, which is a set of hyperlinks pointing to top-level web pages. Usually, the hyperlink pool of a website consists of hyperlinks listed in the site-index page or the site-directory page. Hyperlinks in the portal page of the University of Arizona website are selected from its hyperlink pool. The hyperlink pool consists of hyperlinks in its site-index page. Hyperlinks in the portal page of My Yahoo! are also selected from its hyperlink pool. The pool, in this case, consists of hyperlinks in its site-directory page.

The design of the portal page of a website, which serves as the homepage of a website or a default web portal. They define a new and important research problem ñ hyperlink selection: selecting from a large set of hyperlinks in a given website a limited number of hyperlinks for inclusion in a portal page. The objective of hyperlink selection is to maximize the efficiency, effectiveness and usage of a web site is portal pageWeb pages in a website are organized in a hierarchy in which a high level web page is an aggregation of its low level web pages [Nielson 1999]. For example, the web page of faculty list is one level higher than its corresponding faculty homepages and it is an aggregation of its corresponding faculty homepages. For a university website, top level web pages include the web page of department list and the web page of computing resources etc.

CONCLUSION:

Hyperlink selection is an important but rarely researched problem. It is formally defined the hyperlink selection problem and proposed a heuristic solution method named Link Selector. As a result, hyperlinks selected by Link Selector based on old structure relationships and access relationships could be out-of-date. To keep the selected hyperlinks up-to date, an obvious solution is to re-run Link Selector every time a change occurs. Apparently, for websites with frequent changes, the cost of frequent re-run is unbearable. An efficient solution needs to be developed to monitor both types of changes and to trigger the re-run of Link Selector only when necessary.

Fuzzy clustering

KUSHALA M(172MCA32)

Abstract

Fuzzy clustering is useful clustering technique which partitions the data set in fuzzy partitions and this technique is applicable in many technical applications like crime hot spot detection, tissue differentiation in medical images, software quality prediction etc.

Fuzzy c-means (FCM) is a data clustering technique in which a data set is grouped into *N* clusters with every data point in the dataset belonging to every cluster to a certain degree.

For example, a data point that lies close to the center of a cluster will have a high degree of membership in that cluster, and another data point that lies far away from the center of a cluster will have a low degree of membership to that cluster.

Key Challenges Of Clustering

- *Data*-driven methods
- Selection of distance function (geometry of clusters)
- *Number* of clusters
- *Quality* of clustering results

Applications:

- Bioinformatics
- Image analysis
- Marketing

Bioinformatics:

In the field of bioinformatics, clustering is used for a number of applications. One use is as pattern technique to analyse gene expression data from microarrays or other technology

In this case, genes with similar expression patterns are grouped into the same cluster, and different clusters display distinct, well-separated patterns of expression

Image analysis:

Fuzzy c-means has been a very important tool for image processing in clustering objects in an image. In the 70's, mathematicians introduced the spatial term into the FCM algorithm to improve the accuracy of clustering under noise.FCM algorithms have been used to distinguish between different activities using image-based features

Marketing:

In marketing, customers can be grouped into fuzzy clusters based on their needs, brand choices or other marketing related partitions.

NUMEROSITY REDUCTION IN DATA MINING

MAITHRY A(172MCA33)

Introduction

Data reduction process reduces the size of data and makes it suitable and feasible for analysis. In the reduction process, integrity of the data must be preserved and data volume is reduced. There are many techniques that can be used for data reduction. Numerosity reduction is one of them. Numerosity Reduction is a data reduction technique which replaces the original data by smaller form of data representation. There are two techniques for numerosity reduction- **Parametric** and **Non-Parametric** methods.

Parametric Methods

For parametric methods, data is represented using some model. The model is used to estimate the data, so that only parameters of data are required to be stored, instead of actual data. Regression and Log-Linear methods are used for creating such models.

Regression:

Regression can be a simple linear regression or multiple linear regression. When there is only single independent attribute, such regression model is called simple linear regression and if there are multiple independent attributes, then such regression models are called multiple linear regression. In linear regression, the data are modelled to a fit straight line. For example, a random variable *y* can be modelled as a linear function of another random variable *x* with the equation y = ax+b where a and b (regression coefficients) specifies the slope and y-intercept of the line, respectively. In multiple linear regressions, y will be modelled as a linear function of two or more predictor (independent) variables.

Log-Linear Model:

Log-linear model can be used to estimate the probability of each data point in a multidimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations. This allows a higher-dimensional data space to be constructed from lower-dimensional attributes. Regression and log-linear model can both be used on sparse data, although their application may be limited.

Non-Parametric Methods

Histograms:

Histogram is the data representation in terms of frequency. It uses binning to approximate data distribution and is a popular form of data reduction.

Clustering:

the cluster representation of the data are used to replace the actual data. It also helps to detect outliers in data.

FREQUENT PATTERN MINING OVER DATA STREAMS

Rabia Firdous(172MCA34)

A prefix-tree structure, called CPS-tree (Compact Pattern Stream tree) that efficiently discovers the exact set of recent frequent patterns from high-speed data stream. The CPS-tree introduces the concept of dynamic tree restructuring technique in handling stream data that allows it to achieve highly compact frequency-descending tree structure at runtime and facilitates an efficient FP-growth-based mining technique.

Recently, finding frequent patterns from data streams has become one of the important and challenging problems, since capturing the stream content memory efficiently with a single-pass and efficient mining have been major issues. The FP-growth mining technique is one of the efficient algorithms where the achieved performance gain is mainly based on the highly compact frequency-descending FP-tree structure that ensures the tree to maintain as much prefix sharing as possible. However, the two database scans and prior threshold knowledge requirements of the FP-tree restrict its use in data stream. DS Tree uses the FP growth mining technique to mine exact set of recent frequent patterns from stream data with a single-pass. However, it provides poor compactness in tree structure and inefficient mining phase, since it uses frequency-independent canonical order tree structure. Therefore, novel tree structure, called CPS-tree (Compact Pattern Stream tree), that constructs an FP tree like compact prefix-tree structure with a single-pass over stream data and provide the same mining performance as the FP growth technique through the efficient tree restructuring process.

Comparing different performance issues of our CPS-tree with those of the DS Tree, to find recent frequent patterns from data stream. Runtime includes tree construction, tree restructuring (for CPS-tree only) and mining time. Several real and synthetic datasets are used. In the experiments, the size of window is indicated by the two parameters W and P. The results on memory consumption on different datasets are in the form of the number of nodes. It is clear that the total number of nodes the CPS-tree requires is significantly less compared to that the DS Tree does in each dataset. The reason is that CPS-tree's dynamic tree Restructuring phase enables it to obtain as much prefix sharing as possible that remarkably reduces the number of nodes compared to any frequency-independent tree. In addition, the CPS-tree is free from the 'curse' of 'garbage' nodes. Moreover, it further reduces the size by maintaining only a few tail-nodes compared to ordinary nodes in the tree structure. The runtime comparison between the CPS-tree and DS Tree over datasets of different types has been performed by varying the threshold values and widow parameters. Runtime of both trees for two (one high and one low) min_sup values over different datasets. The runtime distribution for tree construction, tree restructuring (only for CPS-tree), tree update (expired pane deletion time for CPS-tree and 'garbage' node deletion time for DS Tree), mining time (for two min_sup values) and total time are shown explicitly. The data in the table clearly demonstrate that CPS-tree outperforms DS Tree in overall execution time by multiple orders of magnitudes on both high and low min_sup values over all types of datasets used in the experiment, which is due to the remarkable improvement CPS tree achieves in mining time on the dynamically-obtained frequency-descending tree structure. Therefore, from the above experiments we summarize that despite additional tree restructuring cost, our CPS-tree consistently outperforms state-of-the-art algorithms on both runtime and memory consumption in mining exact set of recent frequent patterns from data stream.

CONCLUSION:

The prefix-tree structure CPS-tree that introduces dynamic tree restructuring mechanism in data stream and efficiently finds recent frequent patterns from high-speed data stream with a single-pass.

DOMAIN DRIVEN DATA MINING

SHILPA S(172MCA35)

Domain driven data mining is a data mining methodology for discovering actionable knowledge and deliver actionable insights from complex data and behaviours in a complex environment. It studies the corresponding foundations, frameworks, algorithms, models, architectures, and evaluation systems for actionable knowledge discovery.

Importance of Domain Driven Data Mining

The process of data mining stops at pattern identification. Consequently, a widely seen fact is that

- many algorithms have been designed of which very few are repeatable and executable in the real world
- often many patterns are mined but a major proportion of them are either common sense or of no particular interest to business, and
- end users generally cannot easily understand and take them over for business use.

In summary, we see that the findings are not actionable and lack soft power in solving real-world complex problems. Thorough efforts are essential for promoting the actionability of knowledge discovery in real-world smart decision making. To this end, domain-driven data mining has been proposed to solve and promote the paradigm shift from "data-centred knowledge discovery" to "domain-driven, actionable knowledge delivery." In D^3M, ubiquitous intelligence is incorporated into the mining process and models, and a corresponding problem-solving system is formed as the space for knowledge discovery and delivery.

Applications Of Domain Driven Data Mining

- Real world data mining
- Capital markets
  - Actionable trading agents
  - Actionable trading strategies
- Social security
  - Activity mining
  - Combined mining

DATA MINING IN A NETWORK SECURITY

VEDAVATHI H L (172MCA36)

Introduction:

Network security is any activity designed to protect the usability and integrity of your *network* and data. It includes both hardware and software technologies. Effective *network security* manages access to the network. It targets a variety of threats and stops them from entering or spreading on your network.

Security management

Security management for networks is different for all kinds of situations. A home or small office may only require basic security while large businesses may require high-maintenance and advanced software and hardware to prevent malicious attacks from hacking and spamming. In order to minimize susceptibility to malicious attacks from external threats to the network, corporations often employ tools which carry out network security verifications

How Does Network Security Work?

There are many layers to consider when addressing network security across an organization. Attacks can happen at any layer in the network security layers model, so your network security hardware, software and policies must be designed to address each area. Network security typically consists of three different controls: physical, technical and administrative. Here is a brief description of the different types of network security and how each control works.

Importance of network Security

Organizations and businesses today, need to understand the importance of network security. It doesn't matter whether it is a government organization, a start-up, small-business, or a multi-national, network security should hold the same level of important for them.
since there are different types of attacks that can happen on your computer network.

1. **Interruption:** An interruption attack targets the availability of a DOS or denial of service attack.

2. Interception: This attack is based on gaining unauthorized access to a network. It can be put into action by acquiring valuable or sensitive information.

3. Modification: A modification attack is based on tampering with resources, and will generally change information that is communicated between parties. It could be sending the wrong information to a party in order to cause miscommunication.

THE DARK SIDE: MINING THE DARK WEB FOR CYBER INTELLIGENCE

ZULFIN ARA(172MCA37)

Like any social construct, the Internet has its dark and seedy side. From social media to encrypted chat rooms and the black markets of the deep web, there's a whole world out there lying just beneath the shiny surface of the Internet and it's here that many cyber-attacks are born.

The majority of successful attacks carried out against businesses are preceded by chatter over social media or underground chat forums, depending on the sophistication of the attacker. Many attacks are also openly publicized to earn the attacker kudos. And yet very few businesses are monitoring either legitimate or underground sites for this kind of noise.

This seems nonsensical given that even a very basic level of monitoring can prevent reputational compromise. For example, monitoring search engines and social media sites such as Facebook, github, Google+, LinkedIn, twitter and Reddit can ensure the organization is informed of any compromise of the brand. There are numerous incidents of hackers hijacking slogans or trademarks and impersonating legitimate businesses in order to convince victims the link they are clicking on is bona fide. Monitoring these sites using key words, phrases and search terms is the bare minimum that a business should be doing.

Going beyond that involves the processing of much higher data volumes. In addition to social media posts, there are forums to monitor and videos, and that's just what we can see. Delve a little deeper and you enter the murky but well-established world of the dark web.

Here you'll find Tor IRC anonymizing software, shady versions of social media platforms and community sites, websites called onions as well as black markets such as the notorious Silk Road. There are, for example, dark versions of twitter and Snapchat, dark wikis, and libraries documenting hacks and exploits. 'Onions' usually require some form of invitation or access code and may well be encrypted. Here there's all sorts of data up for grabs and with the right know-how you can scan for data leakages and compromised data such as emails, domain names or intellectual property.

Yet extrapolating this information into a meaningful form that can be used for threat intelligence is no mean feat. The complexity of accessing the dark web combined with the sheer amount of data involved, correlation of events, and interpretation of patterns is an enormous undertaking, particularly when you then consider that time is the determining factor here. Processing needs to be done fast and

in real-time. Algorithms also need to be used which are able to identify and flag threats and vulnerabilities. Therefore, automated event collection and interrogation is required and for that you need the services of a Security Operations Centre (SOC).

The next generation SOC is able to perform this type of processing and detect patterns, from disparate data sources, real-time, historical data etc. These events can then be threat assessed and interpreted by security analysts to determine the level of risk posed to the enterprise. Forewarned, the enterprise can then align resources to reduce the impact of the attack. For instance, in the event of an emerging DoS attack, protection mechanisms can be switched from monitoring to mitigation mode and network capacity adjusted to weather the attack.

But it doesn't stop there. This type of cyber response provides insights into future threats and attacks and that type of advance warning system can provide the business with real insights that can guide future strategy and security spend, helping focus security investment. For example, knowledge of sector specific emerging threats such as spear phishing campaigns can help steer staff training.

The problem to date is that such SOC services have typically been the preserve of big business. Thankfully, the threat intelligence sector is now maturing and the commoditization of services is seeing this kind of deep threat intelligence become available to mainstream business. Tiered SOC services provide entry level options which can scale with the business, ensuring that security budgets are no longer frittered away on numerous point solutions, or the endpoint, but are spent where they are needed.

Going forward, legitimate Internet services will continue to be misapplied and used to attack the enterprise. There's already evidence of spoof social media profiles being created to target high worth individuals, for instance. And the Internet's underbelly, the dark web, will continue to be used to plan and coordinate organized attacks. The issue remains whether organizations will continue to do business in blissful ignorance or arm themselves with this ready stream of threat intelligence. We ignore these dark domains at our peril.

WEB MINING: A KEY TO IMPROVE BUSINESS ON WEB

Divya S(172MCA38)

The web mining concept and how it can be useful and beneficial to the business improvement by facilitating its applications in various areas over the internet. The various areas containing web sites on internet, which can make best use of different web mining techniques to improve their business decisions based on the user behaviour analysis which can ultimately help in improving the relevance of their web site to suit their user needs and adding value to their business growth. It also contributes about the factors responsible and governing the usage of web mining for the web sites to improve business intelligence. The need for techniques which would be able to classify, categories, cluster the web pages in such a way that the web page retrieval can be done in a optimum way and to reduce the burden on the user to keep on searching the required web page from the sea of the information. Web mining is helpful in making business decisions for further trends and patterns of user access of content of the web pages and customer behaviour in an effective way. The major business areas which can be benefited by applying web mining techniques.

STATUS OF WEB INFORMATION

The information on the internet is in the form of static and dynamic web pages of various areas from education, industry to every walk of life including blogs. As per the web sites more than 160,000,000 web sites are having inter, intra linked web pages. The speed of increase of web information is rapid. The hidden knowledge discovery, patterns and trends of user access can be found from the way the web sites and web pages are accessed and it is useful from the business perspective giving future directions for decision making. The Data Mining techniques help in identifying the patterns implying the future trends in the studied data. The Web Mining is an application of the data mining techniques to find interesting and potentially useful knowledge from web data.

The Various Business Areas Where Web Mining has helped in Improving the Business Decision Making

- **E-Business**

Analysis of click-stream data i.e. web mining uncovers real-time e-business opportunities across geography. It provides ways to target right customers and understand their needs and to customize services and strategies in near-or-real time. The area of advertising is no exception for utilizing the opportunities provided by online customer analytics to promote right products in real time to the right customer. It also helps in effectiveness of a web site as a channel for marketing by quantifying the user's behaviour while on the web site.

- **CUSTOMER BEHAVIOUR**

Web Mining helps in understanding the concerns such as current and future probability of every customer, relationship between behaviour and the loyalty at the website The models based on customer-centric web behaviour can be used not only for identifying improvements in the appeal of

web site segmentation, which are based on web behaviour providing a precise basis for personalization but also for predicting customer's future behaviour that is essential for website content planning and design.

- **CRM**

Analytical CRM utilizes business intelligence and reporting methodologies such as data mining and analytical processing to CRM applications. While the earlier CRM implementations focus on improving operational efficiencies in the sales and service functions through tailor-made solutions for call-center management, analytical CRM solutions use intelligence solutions to analyse the data, identify the demographic profiles and measure the purchase frequency and other behavioural patterns of the customers.

With the amount of available online content, today organizations put premium on understanding, adopting and managing the same, convert them into appropriate knowledge suitable to serve their customers better, and thus improve the operations and accelerate the process of delivery of products to markets. The World Wide Web is a fertile area for web Mining and it can provide applications, methods, algorithms to be beneficial in various real-world applications with respect to the critical e-CRM function.

- **CROSS SELLING**

Web Data mining usage which will allow to cross- sell into web store application with a minimal effort.

CONCLUSION

In today's era where the entire world has become a global village and the driving force is internet having e-business to internet blogs to search engines, the major questions in front of the business users is while they would like to retain the existing customers and also would like to understand the patterns and trends of customer behaviour so that their decisions can be supported with facts represented with visualizations and appropriate reporting made possible with web mining. The success of accuracy of deriving patterns is directly proportional to the amount of sample data used for the data mining techniques.

The advantages of using web mining in search engines and e-commerce, CRM, customer behaviour analysis, cross selling; web site service quality improvement is noticeable. The recommendation of using web mining techniques can be applied successfully with a keen analysis of clearly understood business needs and requirements. Also one more governing factor is the amount of data, as the data is voluminous the results can be more towards the correct trends and patterns to be predicted from the given set of data. But although the web mining techniques can be applied to even the small web sites with a few number of web pages and links within them, web mining may not be the answer for its improvement as it will not be the optimum solution as far as the cost factor in terms of parameters such as complexity of web mining techniques using algorithms may not be recommended.

Possible applications can be On-line social networking community software applications can use web mining techniques to explore the effectiveness of on-line networking, also areas such as knowledge management web sites and web mining can also be useful in bioinformatics, e-governance and e-learning.