JYOTI NIVAS COLLEGE

AUTONOMOUS

TECH ON TAP

EJOURNAL: DEPARTMENT OF MCA

ISSUE: 3 **OCTOBER 2017**

INDEX

SL NO	TITLE	PG NO
1	Spatial And Geographic Data Mining	1
2	Multimedia Data Mining	2
3	Issues And Techniques Of Web Mining	4
4	Olap Analysis Operators For Multi-States Data Warehouse	6
5	Search Engine In Data Mining	8
6	Mobility Data Warehousing And Mining	10
7	Data Mining In Cloud Computing	11
8	Data Mining In GenomicsAnd Proteomics	13
9	Efficiently Mining Frequent Tress In Forest	15
10	Big Data & Business Intelligence	16
11	Fraud Detection Using Data Mining	17
12	Survey Of Data Mining Techniques For Social Networking Websites	19
13	Data Mining From Smart Card Data Using Data Clustering	21
14	Data Mining In Health Care	23
15	Educational Data Mining	25
16	Automated Personality Classification Using Data Mining Techniques	27
17	Data Warehousing For Rough Web Caching And Pre-Fetching	29
18	Open Source Data Mining Tools	31
19	Spatial Data Mining	33

SPATIAL AND GEOGRAPHIC DATA MINING

ASMA FARHEEN (152MCA36)

KARIMA.S (152MCA33)

The data types which come to mind when the term data mining is mentioned involves data as we know it statistical, generally numerical data of varying kinds. However, it is also important to consider information which is of an entirely different kind, spatial and geographic data which could contain information about astronomical data, natural resources, or even orbiting satellites and spacecraft which transmit images of earth from out in space. Much of this data is image-oriented, and can represent a great deal of information if properly analyzed and mined.

A definition of spatial data mining is as follows: the extraction of implicit knowledge, spatial relationships, or other patterns not explicitly stored in spatial databases.‖ Some of the components of spatial data which differentiate it from other kinds include distance and topological information, which can be indexed using multidimensional structures, and required special spatial data access methods, together with spatial knowledge representation and data access methods, along with the ability to handle geometric calculations. Analyzing spatial and geographic data include such tasks as understanding and browsing spatial data, uncovering relationships between spatial data items (and also between non-spatial and spatial items), and also analysis using spatial databases and spatial knowledge bases. The applications of these would be useful in such fields as remote sensing, medical imaging, navigation, and related uses. Some of the techniques and data structures which are used when analyzing spatial and related types of data include the use of spatial warehouses, spatial data cubesandspatialOLAP.Spatialdatawarehousescanbedefinedasthosewhicharesubjectoriented, integrated, nonvolatile, and time-variant .

Some of the challenges in constructing a spatial data warehouse include the difficulties of integration of data from heterogeneous sources, and also applying the use of on-line analytical processing which is not only relatively fast, but also offers some forms of flexibility. In general, spatial data cubes, which are components of spatial data warehouses, are designed with three types of dimensions and two types of measures.

The three types of dimensions include the noncapital dimension (data which is noncapital in nature), the spatial to noncapital dimension (primitive level is spatial but higher level-generalization is noncapital), and the spatial-to-spatial dimension (both primitive and higher levels are all spatial). In terms of measures, there are both numerical (numbers only), and spatial (pointers to spatial object) measured used in spatial data cubes. A side from the implementation of data warehouses for spatial data, there is also the issue of analyses which can be done on the data. Some of the analyses which can be done include association analysis, clustering methods, and the mining of raster databases There have been number of studies conducted on spatial data mining.

MULTIMEDIA DATA MINING

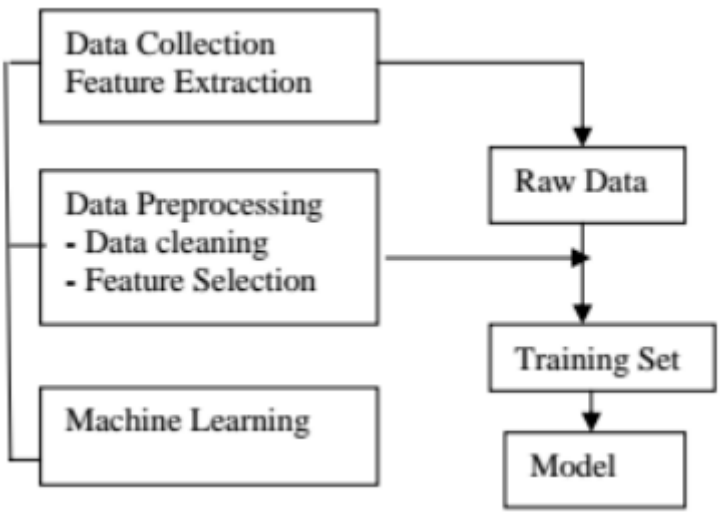
ASHA V(15MCA03)
LOKESHWARI .V(15MCA14)

Introduction

Multimedia data mining refers to the mining of Multimedia content. In other words, it is study of large amounts of multimedia information in order to find patterns or statistical relationships. Once data is collected, computer programs are used to analyze it and look for meaningful connections. Recent progress in the field of electronic imaging, video devices, storage, networking and computer power, show that the amount of multimedia has grown enormously, and data mining has become a One solution is to develop mining tools to operate on the multimedia data directly. The main requisite of Multimedia data mining is the collection of huge amounts of data. The key factor is the sample size when analyzing data because predicted trends and patterns are more likely to be inaccurate with a smaller sample. This data can be collected from various media, including videos, sound files, and images. Some experts also consider spatial data and text to be multimedia

Architectures For Multimedia Data Mining

The Present architecture of applying multimedia mining in different multimedia types . Data collection is the starting point of a learning system, as the quality of raw data determines the overall achievable performance. Then, the goal of data pre-processing is to discover important features from raw data. Data preprocessing includes data cleaning, normalization, transformation, feature selection, etc. Learning can be straightforward, if informative features can be identified at the pre-processing stage. Detailed procedure depends highly on the nature of raw data and problem's domain.



Multimedia mining process

Techniques Of Multimedia Data Mining

1. Multimedia Data Mining Process Using Classification Rules:

In this approach, main focus is on discovering the semantic structures. We use the classification rule approaches to perform data mining process because this approach only induces absolutely accurate rules. Examples of this work are: 1. The Hidden Markov Model 2. Detection of soccer goal shots using decision tree

2.Multimedia Data Mining Process Using Clustering:

Clustering is a process of organizing objects into groups whose members are similar in some way. It is one of the unsupervised learning data mining technique. In unsupervised classification, the problem is

to group a given collection of unlabeled multimedia files into meaningful clusters according to the multimedia content without Apriority knowledge.

3.Multimedia Data Mining Process Using Association Rules:

For discovering interesting relations between variables in large databases, the Association rule learning is a popular and well researched method. There are different types of associations which are association between image content and non-image content features.

Some early examples are: 1. Image classification method by using multiple level association rules based on image objects. 2. A multi relational extension to FP-tree algorithm to accomplish association rule mining task effectively.

Applications Of Multimedia Data Mining

Multimedia data mining for traffic video sequences – Example: Traffic camera footage to analyze traffic flow. This would come in handy while planning new streets, expanding existing streets, or diverting traffic. The same can be used by the Government organizations and city planners to help traffic flow more smoothly and quickly.

Multimedia Data Mining in Digital Libraries — The Digital library retrieves, stores and preserves the digital data. For this purpose, there is a need to convert different formats of information such as text, images, video, audio, etc. Thus, in the process of conversion of the multimedia files in the libraries, the data mining techniques are popular.

Application in medical analysis - Application of Data Mining Techniques for Medical Image Classification Media Production and Broadcasting – Proliferation of radio stations and TV channels makes broadcasting companies to search for more resourceful methodologies for creating programs and monitoring their content.

Future Scope

There are several future opportunities in the area of large-scale retrieval and mining that are worthy of attention from the multimedia community:

1 Many technical issues are yet to be addressed when managing large multimedia collections, for example, obtaining accurate annotations, efficient indexing of visual features, best possible creation of large-scale benchmark collections, and organizing these annotations.

2 Most state-of-the-art, machine-learning algorithms, such as nonlinear kernel support vector machines, kernel logistic regression, and k-means, can't be easily extended to large collections because their computational complexities are quadratic or even cubic with the size of the training set.

3User interface, visualization, and interaction patterns will become more complicated with large quantities of data.

4.Distributed- and cloud-computing platforms as well as parallel machine-learning and datamining algorithms will become necessary to make large-scale multimedia analysis run at practical speeds.

Conclusion

This paper proposes a survey of multimedia data mining. The key idea is to provide review of Multimedia Data Mining, which is an active and growing area of research. While the majority of the work has been devoted to the development of data mining methodologies to deal with the specific issues of multimedia data, several applications of multimedia data mining have been investigated.

ISSUES AND TECHNIQUES OF WEB MINING

KUMARI POONAM(15MCA12)
SRISHTI PAL (15MCA11)

Introduction

Web mining is the application of data mining technique which is an unstructured or semi-structured data. It automatically discovers and extracts potentially useful and previously unknown information or knowledge from the web. The significant web mining application are Web design, web search engines,informationretrieval,networkmanagement,E-commerce,business and artificial intelligence ,webmarket places and web communities. Basically it is the application of data mining techniques to discover patterns from the web.

Web mining can be divided into three different types:-
Web usage mining : 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. Web usage mining can also referred as automatic discovery and analysis of patterns. Web usage mining itself can be classified further depending on the kind of usage data considered.

Web Server Data: The user logs are collected by the web server. Typical data includes IP address, page reference and access time.

Application Server Data: Commercial application servers have significant features to enable ecommerce applications to be built on top of them little effort.

Application Level Data: New kinds of events can be defined in an application, and logging can be turned on. Generating histories is the spatiality of the events.

Here are the four techniques of web usage mining:

Sequential pattern mining based: Allows the discovery of temporally ordered Web access patterns.
Association rule mining based: Find correlation among Web pages.
Clustering based: Groups users with similar characteristics.
Classification based: Groups users into predefined classes based on their characteristics.
Web structure mining

Web structure mining is the study of data interconnected to the structure of a particular website.It consists of web graph which contains the web pages or web documents as nodes and hyperlinks as edges those are connecting between two related pages.Web structure mining can be performed either at intra-page level orinter-page level.Ahyperlink that connects to a different part of the same page is called intra-page hyperlink.It is a document structure level.A hyperlink that connects two different pages are called inter-page hyperlink which is structure level.Web page is organized in tree structure format based on HTML tags.Here,the documents are extracted automatically by the Document Object Model(DOM).Web structure mining is used in serach engines such as Google,Yahoo etc.
Issues on Web Structure Mining
Web structure mining has two issues due to its huge amount of data.
Reducing irrelevant search results.Relevance of search information becomes unorganized due to the problem search engines oftenonly tolerate for low precision criteria.
Indexing information on the web.This causes low amount of recall with content mining.

Content Mining

Web content mining data may be structured or unstructured/semi structured even though much of web is unstructured.It is the process of retrieving the information from the web into more structured forms and indexing the information to retrieve quickly or finding valuable information from web content or web documents. Web content miming used many algorithms and tools such as Genetic algorithm,Cluster Hierarchy Construction Alogorithm (CHCA),Correlation algorithm.It has two approaches:

Agent Based Approach

Agent based approach focuses on searching relevant information from the World Wide.These types of agents are

Intelligent search agents: Automatically searches for information along with a particular query.

Information filtering /categorizing agents: Filters the data personalized web agents Discovers the documents those are related to the user profiles.

Database Approach

Database approach consists of databases which contain attributes,tables and schema with defined domains.It focused on techniques for organizing the semi structured data on the web into more collections of resources,and using standard database querying mechanism and data mining techniques to analyze it.

Issues on Web Content Mining

Web content mining has number of resarch issues because it can extract the information from the web search engines.

Data/Information Extraction concentrate on extraction of structured data from web pages such as products and search results.

Web information integration and schema matching.

Conclusion

In this focus on representation issues,various techniques of web usage mining and web structure mining and information retrieval and extraction issues in web content mining and connection between the web content mining and structure mining.

OLAP ANALYSIS OPERATORS FOR MULTI-STATES DATA WAREHOUSE

VINOLIYA GRACE.E (152MCA35)
NANCY ESTHER.D (152MCA34)

Introduction

Multidimensional Data Warehouse (MDW) organizes data in a multidimensional way in order to support On-Line Analytical Processing (OLAP). More specifically, a MDW schema is based on facts (analysis subjects) and dimensions (analysis axes). The facts contain analysis indicators, while a dimension includes analysis parameters organized in hierarchies from the minimal (most detailed) granularity to the maximal granularity.

In a classical MDW, all data are permanently stored and new data is periodically added. The increasing volume of MDW makes the tasks of decision-makers more difficult since they may be lost during their analyses. On the other hand, information is usually timely sensitive; most of detailed information loses its value over time. Nevertheless, information at high granularity levels is more stable, and it can generally fulfil decision-makers needs when analyses are carried out over older data. For instance, an analyst may have interest in analysing sale amounts by product's brand for the last five years. However, as lots of today's brands did not exist before, the brand granularity level may be useless for an older period. As a result, the analyst may have no more interest in analysing sale amounts by brand over the last ten years but by a higher and more stable granularity level, such as product's category.

Facing large volumes of data with a great amount of inadequate detailed data causes inefficiency in analysis. Therefore to eliminate the inadequate detailed data we use the Multi-Dimensional Data Warehousing (MDW) model which is based on data reduction to aggregate and then remove useless detailed data. The OLAP operator supports analyses over reduced data.

Multi-States Analysis Operators

A reduced MDW facilitates decision-makers tasks by keeping only useful data over time. In order to carry out analyses over reduced data, a decision-maker needs OLAP operators applicable to MDW composed of multiple states of user-oriented OLAP operators supporting displaying, drilling, rotating and selecting operations.

There are five commonly-used OLAP operators in order to support basic multi-states analysis in reduced MDW.

Display multi-states: It allows building the first analysis results.

Drilldown multi-states: They allow decreasing analysis granularity.

Rollup multi-states: They allow increasing analysis granularity.

Rotate multi-states: These operators allow changing the content of a currently displayed analysis axis.

Select multi-states: Allows adding restriction predicates to analysis results.

The operators can be easily implemented in different environments (e.g. ROLAP, MOLAP and HOLAP).

Characteristics

User-oriented algebraic operators:

Each analysis operator is user-oriented. It defines an elementary analysis operation in the point of view of a decision-maker. To facilitate decision-makers tasks, the multi-states analysis operators manipulate conceptual concepts such as fact and dimensions in the context of reduced MDW.

Bi-directional adaptation:

Bi-directional adaptation Analysis cannot be carried out with missing elements (i.e. missing attributes, missing hierarchies and missing dimensions). Adaptations should be taken if missing elements are involved during an analysis.

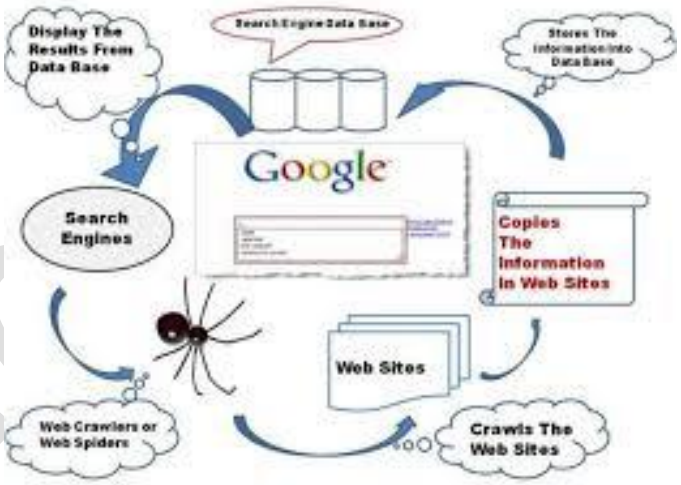
CONCLUSION

With the use of Multi-Dimensional Data Warehousing model we can eliminate the inadequate data which is based on the data reduction which removes useless data and with the use of OLAP Analysis operators one can efficiently analyse the reduced data.

SEARCH ENGINE IN DATA MINING

H SRINAVYA B (15MCA06)
RAKSHA (15MC20)

Search engine analysis takes place by regularly monitoring huge volumes of data arising from internet usage statistics, keyword usage statistics and many other parameters. Data mining tools can keep track of these data by efficiently storing them, analysis them and producing output as and when necessary.
When we use the term search engine in relation to the internet they are usually referring to the actual search forms that search through databases of HTML documents available all over the internet. Internet is an immense, huge and dynamic data collection that includes infinite hyperlinks and volumes of data usage information – hence requires effective data mining. But huge data is still a challenge in knowledge discovery.
Web based search engine works by saving the information of many web pages, which they retrieve itself. These pages are retrieved by a web crawler which is also called spider which follows every link on the site. Search engine is a term used for information retrieval search engine match queries against an index that they create. This index contains the word in each document, pointers to their location within the document. This is called inverted file.
Working of search engine:
There are basically three types of search engine those are powered by crawlers, ants or spiders and those that are powered by humans and those that are a combination of two.



Crawlers based search engine are those that use automated software (called crawlers) that visit a web site, read the information on the actual web site, read the site's, meta tags and also follow the links that the site links to performing indexing on all linked web sites as well. The crawler returns back all that information to a central depository, where the data is stored and index the crawler will periodically return to the sites to check for any information that has updated. The frequency with which this happens is determined by the administration of the search engine and it also affects the efficiency of the search engine.
Human powered search engines depend on humans to submit information that is subsequently indexed and catalogued. Only information that is submitted is put into the index.
This search engine in data mining approach can increase the raking of a website and benefit its owners and provide users with more accurate and relevant search results.

MOBILITY DATA WAREHOUSING AND MINING

SHINY ELIZABETH.S(15MCA26)
SAI JYOTHI(15MCA30)

The flow of data generated from low-cost modern sensing technologies and wireless telecommunication devices enables novel research fields related to the management of this new kind of data and the implementation of appropriate analytics for knowledge extraction.The analysis of such mobility data raises opportunities for discovering behavioral patterns that can be exploited in applications like mobile marketing, traffic management etc.

Online analytical processing (OLAP) and data mining (DM) techniques can be employed in order to convert this vast amount of raw data into useful knowledge. Their application on conventional data has been extensively studied during the last decade. The high volume of generated mobility data arises the challenge of applying analytical techniques on such data. In order to achieve this aim, we have to take into consideration the complex nature of spatiotemporal data and thus to extend appropriately the two aforementioned techniques to handle them in an efficient way.

Towards this direction, we provide two motivation scenarios. Firstly, let us consider an advertising company which is interested in analyzing mobility data in different areas of a city so as to decide upon road advertisements (placed on panels on the roads). They are interested in analyzing the demographical profiles of the people visiting different urban areas of the city at different time zones of the day so as to decide about the proper sequence of advertisements that will appear on the panels at different time periods. This knowledge will enable them to execute more focused marketing campaigns and apply a more effective strategy.

Indicatively, a Trajectory Data Warehouse (TDW) can serve this aim by analyzing various measures such as the number of moving objects in different urban areas, the average speed of vehicles, the ups and downs of vehicles' speed as well as useful insights, like discovering popular movements.

Secondly, trying to understand, manage and predict the traffic phenomenon in a city is both interesting and useful. For instance, city authorities, by studying the traffic flow, would be able to improve traffic conditions, to react effectively in case of some traffic problems and to arrange the construction of new roads, the extension of existing ones, and the placement of traffic lights.

The above targets can be served by analyzing traffic data so as to monitor the traffic flow and thus to discover traffic related patterns. These patterns can be expressed through relationships among the road segments of the city network. In other words, we aim to discover, by using aggregated mobility data, how the traffic flows in this network, the road segments that contribute to the flow and how this happens.

In order to realize the two above scenarios, but also many others, we work on a framework for Mobility Data Warehousing and Mining that takes into consideration the complete flow of tasks required for the development of a TDW and the application of trajectory-inspired mining algorithms so as to extract traffic patterns.

Warehousing Spatial And Mobility Data

The pioneering (SDW). The authors extend the idea of cube dimensions so as to include spatial and non-spatial ones, and of cube measures so as to represent space regions and/or calculate

Trajectory warehousing is in its infancy but we can distinguish three major research directions on this field: modeling, aggregation and indexing. From a modeling perspective, the definition of hierarchies in the spatial work by Han et al. introduces the concept of spatial data warehousing dimension introduces issues that should be addressed. The spatial dimension may include not explicitly defined hierarchies. Thus, multiple aggregation paths are possible and they should be taken into consideration during OLAP operations. Tao and Papadias propose the integration of spatial and temporal dimensions and present appropriate data structures that integrate spatiotemporal indexing with pre-aggregation. Choi et al. try to overcome the limitations of multi-tree structures by introducing a new

index structure that combines the benefits of Quadtrees and Grid files. However, the above frameworks focus on calculating simple measures (e.g. count customers).

Furthermore, an attempt to model and maintain a TDW is presented in where a simple data cube consisting of spatial / temporal dimensions and numeric measures concerning trajectories is defined. In our research, we investigate efficient solutions to support complex measures and to define the complete flow of processes in a TDW.

Mining Patterns From Mobility Data

In a distributed traffic stream mining system is proposed: the central server performs the mining tasks and ships the discovered patterns back to the sensors, whereas the sensors monitor whether the incoming traffic violates the patterns extracted from the historical data. This work emphasizes on the description of the distributed traffic stream system, rather on the discovery of traffic related patterns.

Also, relative to our research is the work by for the discovery of hot routes (sequences of road segments with heavy traffic) in a road network. The authors propose a density-based algorithm, called FlowScan, which cluster road segments based on the density of the common traffic they share. The algorithm, however, requires the trajectories of the objects that move within the network, thus cannot be applied in our problem settings (as we already mentioned and will be further explained in Section 3.4, we assume aggregated mobility data and not the trajectories of each object).

A line of research relevant to our work is that of *spatiotemporal* or *trajectory clustering* that aims at grouping trajectories of moving objects into groups of similar trajectories.

Lee et al propose a partition-and-group framework for trajectory clustering. Similar line segments are grouped into a cluster using a density based clustering method. For each cluster, the representative trajectory is discovered which is defined as the trajectory describing the overall movement of the trajectory partitions that belong to the same cluster. This work concerns the trajectories of the moving objects, free movement and no some predefined network like, in our case, the road network.

Giannoti et al.propose the notion of trajectory patterns (Tpatterns) and introduce appropriate trajectory mining algorithms for their discovery. Trajectory patterns represent sequences of spatial areas of interest that are temporally related. Such areas of interest can be predefined by the user or they can be discovered in a dynamic way using some density-based algorithm.

Kalnis et al.introduce the notion of moving clusters for discovering groups of objects that move close to each other for a long time interval. However, their method requires the IDs of the objects and considers unconstrained environments.

DATA MINING IN CLOUD COMPUTING

ANGEL CHRISTINA (152MCA31)
SOUMYA.P.H (15MCA27)

Data Mining is a process of extracting potentially useful information from raw data. Cloud Computing denotes the new trend in internet services that rely on clouds of servers to handle tasks. Data mining in cloud computing is the process of extracting structured information from unstructured or semi-structured web data sources. The data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable and secure services for their users

The data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable and secure services for their users. The implementation of data mining techniques through Cloud Computing will permit the users to get back significant information from virtually integrated data warehouse that reduces the cost of infrastructure and storage. CDM (Cloud Data Mining) offers fabulous potential for analyzing and extracting the information in various fields of human activities: business, economics, health care medicines, heredity, biology, pharmacy, advertising, etc. Cloud provides tools that can "handle" large volume of data, which cannot be processed efficiently and at reasonable cost using certain techniques.

The main effects of data mining tools being delivered by the Cloud are:

The customer only pays for the data mining tools that he needs – that reduces his costs since he doesn't have to pay for complex data mining suites that he is not using exhaustive;

The customer doesn't have to maintain a hardware infrastructure, as he can apply data mining through a browser – this means that he has to pay only the costs that are generated by using Cloud computing.

Data Mining in Cloud:

Association Rule: An association rule mining helps in finding relation between the items or item sets in the given data. The association rule mining algorithm in Hadoop evaluates testing it in the cloud (EC2) by increasing the number of nodes in the testing set up. The input data is divided among the nodes. Further, the data transfer among the nodes and the situations like a storage node dies or, what would happen if some nodes in the cluster does not run, are taken care by Hadoop. This adds a great deal of robustness and scalability to the system.

Apriori algorithm is used in the cloud computing environment to solve traditional problems encountered in the traditional Apriori data mining. Apriori Algorithm Research achieved high extension capability based on Hadoop platform, and proves the possibility of association rule mining algorithm and cloud computing technologies. To gain more experience in cloud-assisted data mining, the association rule based algorithm, Apriori proved how data mining algorithms can be adjusted to fit the increasing demand for parallel computing environment of cloud.

Classification, Regression, Summarization: We can use cloud computing for fingerprint data storage and recognition. There are several areas we can visualize, where data mining can be applied. A particular data mining algorithm is usually an instantiation of the model-preference-search components. The models that can be useful for fingerprint data handling are: Classification, Regression, Clustering and Summarization. Cloud computing supplies cheap and efficient remedies for storing and analyzing mass data.

Clustering: Clustering model for assessing SaaS will help to evaluate possible software services on the cloud computing by using Data Mining Clustering algorithms. The clustering model would be highly useful to software service providers to evaluate their own services to the cloud users. It helps service provider to increase availability of software services on the cloud computing environment

suitable for cloud users needs. It also helps cloud users to evaluate potential software services available on the cloud computing environment.

Prediction: In resource usage prediction algorithm uses a set of historic data to recognize similar usage patterns to a current window of records that occurred in the past. The algorithm then predicts the system usage by interpolating what follows after the identified patterns from the historical data.

Data mining technologies provided through Cloud computing is an essential characteristic for present day businesses to make proactive, knowledge driven decisions, as it helps them have future trends and behaviorspredicted.

DATA MINING IN GENOMICSAND PROTEOMICS

BABITHA K(152MCA38)
SHELOMITH FERNANDES**(**152MCA37**)**

Journal of Data mining in Genomics and Proteomics is one of the best Open Access journals of Scholarly publishing that aims to publish the most complete and reliable source of information on the discoveries and current developments in the mode of original articles, review articles, case reports, short communications, etc. in all areas of the field and making them freely available through online without any restrictions or any other subscriptions to researchers worldwide.

Genomics maps the genetic structure of the living organism, Proteomics explores the ways and means to utilize the molecular biology, biochemistry and genetics to analyze the structure, function and interaction of the proteins, produced by the genes. Data mining is a tool used to extract precious, yet unknown information from huge database to transform the data into useful information.

Data Mining Journals are at higher echelons that enhance the intelligence and information dissemination on topics closely related to data mining. They provide a unique forum dedicated to scientists to express their research articles, review articles, case reports and short communications on an array of data mining research. The Data Mining Peer Reviewed Journals are proficiently supported by universally prominent Editorial Board members.

Genomic data warehousing

A number of generic data warehousing frameworks have been developed to facilitate the integration and querying of genomics data. Some of genomic data warehouses are BioMart, BioXRT, InterMine and PathwayTools.

Genomic data mining

Advances in various high-throughput biotechnologies such as RNA gene expression microarrays. genomic data mining approaches have been successful and represent a promising direction for future work

Data mining applications in genomics

Data Mining and Applications in Genomics contains the data mining algorithms and their applications in genomics, with frontier case studies based on the recent and current works. It provides a systematic introduction to the use of data mining algorithms as an investigative tool for applications in genomics.

Data mining applications in proteomics

Data Mining applications in Proteomics from Standards to Applications, experts in the field present these new insights within the proteomics community, taking the historical evolution as well as the most important international standardization projects into account. An enormous amount of data was created, leading to a wide-spread rethinking of strategy design and data interpretation.

Proteomics data warehousing

The extreme complexity of the Proteome calls for different multistep approaches for separation and analysis on protein and on peptide level. These are usually combinations of 1D or 2D gel electrophoresis and one- to multidimensional LC techniques in combination

with different MS and MS/ MS techniques, all of which are supported by the ProteinScapedata warehousing concept.

Data algorithms

It is a set of heuristics and calculations that creates a data mining model from data. To create a model, the algorithm first analyzes the data we provide, looking for specific types of patterns or trends.

Data modelling and intellegence

Data modeling is often the first step in database design and object-oriented programming as the designers first create a conceptual modelof how data items relate to each other. Data modeling involves a progression from conceptual model to logical model to physical schema.

Data mining tools

Microsoft SQL Server Analysis Services provides so many tools that you can use to create data mining solutions. The Data Mining Wizard in SQL Server Data Tools (SSDT) makes it easy to create mining structures and data mining models, using either relational data sources or multidimensional data in cubes.

Proteogenomics

Proteogenomics has emerged as a field at the ju nction of genomics and proteomics. It is a loose collection of technologies that allow the search of tandem mass spectra against genomic databases to identify and characterize protein-coding genes.is an emerging field of biological research at the intersection of proteomics and genomics.

Computational drug design

Computational methods to predict modes of protein-protein interaction, as well as protein interface hot spots, have garnered significant interest, in order to facilitate the development of drugs to successfully disrupt and inhibit protein-protein interactions.

EFFICIENTLY MINING FREQUENT TRESS IN FOREST

PRIYADHARSHINI(15MCA19)
T.MARY CHROSINE (15MCA29)

Frequent Structure Mining (FSM) refers to an important class of exploratory mining tasks, namely those dealing with extracting patterns in massive databases representing complex interactions between entities. FSM not only encompasses mining techniques like associations and sequences [but it also generalizes to more complex patterns like frequent trees and graphs .Such patterns typically arise in applications like bioinformatics, web mining, mining semi structured documents, and so on. As one increases the complexity of the structures to be discovered, one extracts more informative patterns; we are specifically interested in mining tree-like patterns.

Goal:

*) To efficiently enumerate all frequent sub trees in a forest (database of trees) according to a given minimum support (*minsup*)

*) The support of a sub tree S is the number of trees in D that contains one occurrence of S.

*) A sub tree S is frequent if its support is more than or equal to a user specified *minsup* value.

In this paper we introduce TreeMiner, an efficient algorithm for the problem of mining frequent sub trees in a forest (the database). The key contributions of our work are as follows:

1) We introduce the problem of mining embedded sub trees in a collection of rooted, ordered, and labeled trees.

2) We use the notion of a scope for a node in a tree. We show how any tree can be represented as a list of its node scopes, in a novel vertical format called scope-list.

3) We develop a framework for non-redundant candidate sub tree generation, i.e., we propose a systematic search of the possibly frequent sub trees, such that no pattern is generated more than once.

4) We show how one can efficiently compute the frequency of a candidate tree by joining the scope-lists of its sub trees.

5) Our formulation allows one to discover all sub trees in a forest, as well as all sub trees in a single large tree. Furthermore, simple modifications also allow us to mine unlabeled sub trees, unordered sub trees and also frequent sub-forests (i.e., disconnected sub trees).

Applications:

Mining frequent trees is very useful in domains like bioinformatics, web mining, mining semi-structured data, and so on. We formulate the problem of mining (embedded) sub trees in a forest of rooted, labeled, and ordered trees. We use TreeMiner, a novel algorithm to discover all frequent sub trees in a forest, using a new data structure called scope-list. We contrast TreeMiner with a pattern matching tree mining algorithm (Pattern Matcher)

Conclusion:

Introduce the notion of mining embedded sub trees in a (forest) database of trees

Systematic candidate sub tree generation. No sub tree is generated more than once. (but has a mistake)

Use a string encoding of tree to store dataset efficiently

Use a node's scope to develop scope-lists

Introduce a new algorithm – TreeMiner.

BIG DATA & BUSINESS INTELLIGENCE

NEETU KUMARI **(**15MCA10)
ANSHU KUMARI **(**15MCA02)

Big data is the current buzz word in the industry, both commercial and scientific, that is driving everyone crazy. As the internet is blooming with IOT, data are generated in huge volumes and there rising concerns about its storage, processing and usage.
The impact of big data is undeniable; newspapers and academic journals are full of anecdotes and case studies that illustrate the value of such data to businesses.
Analytics and business intelligence are clearly related, however extracting business intelligence from big data is not as straightforward as it might seem.
Integrating advanced analytics for big data with BI systems is an important step toward gaining full return on investment. Advanced analytics and BI can be highly complementary; advanced analytics can provide the deeper, exploratory perspective on the data, while BI systems provide a more structured user experience. BI systems' richness in dashboard visualization, reporting, performance management metrics, and more can be vital to making advanced analytics actionable.

Special Issue: Big Data Analytics for Business Intelligence
Big data, characterized with high volume, variety and velocity, increasingly drives decision making and is changing the landscape of business intelligence, from governments, organizations, communities to individual decision making. Big data analytics that discover insights from evidences has a high demand for computing efficiency, knowledge discovery, problem solving, and event prediction/prescription. It also poses great challenges in terms of data, process, analytical modeling and management for organizations to turn big data into big insight.

Issue on Big Medical/Healthcare Data Analytics:
Healthcare organizations need to continuously discover useful and actionable knowledge and gain insight from raw data for various purposes such as saving lives, reducing medical errors, increasing efficiency, reducing costs and improving patient outcome. In addition, with the world's population increasing and everyone living longer, models of treatment delivery are rapidly changing, and many of the decisions behind those changes are being driven by data. It becomes crucial to understand as much and as early as possible every patient, hopefully picking up warning signs of serious illness at an early enough stage that treatment is far simpler, and less expensive, than if it had not been spotted until later.

FRAUD DETECTION USING DATA MINING

AMALA JACOB(15MCA01)
MAYA A U(15MCA16)

Today, telecommunication market all over the world is facing a severe loss of revenue due to fierce competition and loss of income due to fraud. To survive in the market, telecom operators usually offer a variety of data mining techniques for fraud detection. According to telecom market, the process of subscribers (either prepaid or postpaid) fraud continues to happen for any telecom industry, it would lead to the great loss of revenue to the company. . In this situation, the only remedy to overcome such business hazards and to retain in the market, operators are forced to look for alternative ways of using data mining techniques and statistical tools to identify the cause in advance and to take immediate efforts in response. This is possible if the past history of the customers is analyzed systematically.

* Predictive Modeling—the process of taking patterns discovered from the database and using them to predict the future.

*Forensic Analysis—the process of applying the extracted patterns to find anomalous or unusual data elements.

Types of Fraud

The types of frauds include credit card frauds, telecommunication frauds, and computer intrusion.

Credit Card Fraud: Credit card fraud is divided into two types: offline fraud and online fraud. Offline fraud is committed by using a stolen physical card at storefront or call center.

Computer Intrusion: Defined as the potential possibility of a deliberate unauthorized attempt to access information, manipulate information, or render a system unreliable or unusable. Intruders may be from an outsider (or hacker) and an insider who knows the layout of the system, where the valuable data is and what security precautions are in place.

Telecommunication Fraud: Fraud is costly to a network carrier both in terms of lost income and wasted capacity. The various types of telecommunication fraud can be classified into two categories: subscription fraud and superimposed fraud. Subscription fraud occurs from obtaining a subscription to a service, often with false identity details, with no intention of paying. Cases of bad debt are also included in this category. Superimposed fraud occurs from using a service without having the necessary authority detected by the appearance of unknown calls on a bill.

Data Mining Techniques In Fraud Detection

Credit Card Fraud Detection:

Credit card fraud detection is quite confidential and is not much disclosed in public. Some available techniques are discussed as follows.

Outlier Detection: An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. Unsupervised learning is a new explanation or representation of the observation data, which will then lead to improved future responses or decisions.

Neural Networks. A neural network is a set of interconnected nodes designed to imitate the functioning of the human brain. Each node has a weighted connection to several other nodes in adjacent layers. Individual nodes take the input received from connected nodes and use the weights together with a simple function to compute output values, neural networks can be constructed for supervised.

Computer Intrusion Detection: Intrusion detection approaches can be broadly classified into two categories based on model of intrusions: misuse and anomaly detection. The techniques used in misuse detection and anomaly detection are described as follows:

Expert Systems. An expert system is defined as a computing system capable of representing and reasoning about some knowledge-rich domain with a view to solving problems and giving advice.

Model-based Reasoning. Model-based detection is a misuse detection technique that detects attacks through observable activities that infer an attack signature. Garvey and Lunt combined models of misuse with evidential reasoning.

Telecommunication Fraud Detection:

Most techniques use Call Detail Record data to create behavior profiles for the customer, and detect deviations from these profiles. These approaches are discussed as follows.

Rule-based Approach: A combination of absolute and differential usage is verified against certain rules in the rule based approach mapped to data in toll tickets.

Visualization Methods: Visualization techniques rely on human pattern recognition to detect anomalies and are provided with close-to-real-time data feeds.

SURVEY OF DATA MINING TECHNIQUES FOR SOCIAL NETWORKING WEBSITES

KAVYASHREE (15MCA09)
SHINY D (15MCA25)

Introduction

Social network has gained remarkable attention in the recent decade. Social network sites such as Twitter, Facebook, accessing them through the internet and the web 2.0 technologies has become more comfortable. People are more interested in and relying on social network for news, information and opinion of other users on diverse subject matters. This often makes social network data complex to analyse manually and results in the pertinent use of computational means to analyse them. Various data mining are used for detecting useful knowledge from massive datasets like trends, patterns and rules. These techniques were used in information retrieval, statistical modelling and machine learning.

Social network is a graph containing nodes and links which is used for representing the social relations on social network websites. A node includes many entities and the relationships between them forming links.

Methods:

Graph Based K-Means Clustering

Graph theory is probably the main method in social network analysis in the early history of the social network concept. K-Means algorithm is the simplest and most commonly used vector quantisation method. K-means clustering partitions data into clusters and minimises distance between cluster centers and data related to clusters. The approach is applied to social network analysis in order to determine important features of the network such as the nodes and links.

Aspect-Based/Feature-Based Opinion Mining

Aspect-based also known as feature-based analysis is the process of mining the area of entity customers has reviewed. This is because not all aspects/features of an entity are often reviewed by customers. It is then necessary to summarise the aspects reviewed to determine the split of the overall review whether they are positive or negative. The sentiments expressed on some entities are easier to analyse than others, one of the reason being that some reviews are ambiguous. The aspect-based opinion problem lies more in blogs and forum discussions than in product or service reviews. The aspect/entity (which may be a computer device) review is either ‘thumb up’ or ‘thumb down’, thumb up life form positive review while thumb down means review negative. Conversely, in blogs and forum discussions both aspects and entity are not familiar and there are high levels of insignificant data which constitute noise. It is therefore necessary to identify opinion sentences in each review to determine if indeed each opinion sentence is positive or negative. Opinion sentences can be used to summarize aspect-based opinion which enhances the overall mining of product or service review.

Homophily Clustering In Opinion Formation

One way to find communities is to use the principle of homophily, which mean that two people tend to communicate more often if they share similar views. Using this phenomenon, those bloggers who have more edges within themselves can be considered as a community. Identifying a set of bloggers that communicate more often among them implies that they share similar views, opinions, or interest; hence they form a community. However, this

approach to community discovery is purely based on network information. Opinion extraction identifies subjective sentences with sentimental classification of whichever positive or negative.

Sentiment Analysis of Social Network

Sentiment analysis also called as opinion mining. The main aim of it is to define the automatic tools able to extract one-sided information from texts in natural language, such as opinions and sentiments, so as to create structured and actionable knowledge to be used by either a decision support system or a decision maker. Sentiment analysis can be referred to as discovery and recognition of positive or negative expression of opinion by people on diverse subject matters of interest. Diverse algorithms are in use to ascertain sentiment that matters to a topic, text, document or personality under review. The purpose of sentiment analysis on social network is to recognize potential glide in the society as it concerns the attitudes, observations, and the expectations of stakeholder or the populace. This recognition enables the entities concern to take prompt actions by making necessary decisions. It is important to decode sentiment expressed to useful knowledge by way of mining and analysis.

Topic Detection and Tracking on Social Network

Topic Detection and Tracking (TDT) on social network employs different techniques for discovering the emergent of new topics (or events) and for tracking their subsequent evolvements over a period of time. TDT is delivery high level of attention recently. Many researchers and authors are conduct research on TDT on social network sites, especially on Twitter. The main aim of TDT was to develop core technologies for news understanding systems. More specifically its tasks focused on discovering and keeping track of real world events in multi lingual news streams from various sources. Various methods have been residential for this task, including machine learning and query expansion based methods.

DATA MINING FROM SMART CARD DATA USING DATA CLUSTERING

JUKUR MEGHA RANI(152MCA32)
SHWETHA T G (152MCA39)

The aim of this paper is to develop an effective methodology for the better understanding of the travelling patterns and evaluating behavioral attributes of traveler's trip. Using smart card data, the data such as boarding location, boarding time, alighting location and alighting time of the traveler is collected and through this data, behavior of the traveler is understood. Methodologies used are pattern recognition and data clustering. This implementation would facilitate the transit authorities to improve the transport service.

More or less 74% of those studying in the university uses bus facility to reach their destination as the university spread 350 Acers of area building are construct in vast area hostellers use means of bus and information gathered for the daily travel This example is presently getting to be evident in creating best university, for example, University use a lot of student depend on bus facility. for instance, more than 32% of day scholars uses their own vehicle Open travel has long been considered to give a powerful approach to decrease blockage, air contamination, and vitality utilization To enhance travel most frequently of bus and sway more individuals to utilize open travel, travel offices have been striving to recognize the key components that draw in travel riders through mulling over their travel designs.

Data description

University has provided smart cards in form of ID cards to their students that could be used for tracking the student and improving the transit services. Transit riders are required to swipe their smart cards when checking-in and checking-out. They have to hold their smart cards near the card reader device to complete the process of entering or exiting buses. Due to which information regarding the boarding and alighting locations including the boarding and alighting time is stored.

Methodology

The two main objectives of this study travel patterns are pattern recognition and regularity mining. A flow diagram of the work performed for the study is illustrated :

(1) Extracting per day smart card data of each sitter from the database.

(2) Using space and time relationship extract passenger's trip chain.

(3) Extracting the passenger's travel pattern and travel regularity based on the generated trip chains by applying sequential data mining approach.

Comparison of data mining Algorithms

Practical passage bundles for student see how student practices are prone to change because of a few charge structures, and accordingly select a passage strategy that attains the ideal harmony between upgrading the engaging quality of the student framework and boosting passage income. University organizers and explores can likewise use individual student-conduct information for action based outing displaying and travel interest breaks down. Data on the travel designs for individual student can likewise be used to evaluate the adequacy of distance travel. Specifically, individual student conduct information can uncover how TOD inhabitants change their everyday driving practices and how student use shifts spatially and transiently. In any case, getting individual record customary travel design examination to a great extent depends on rider fulfilment reviews or travel journals, which is unreasonable and troublesome to actualize at a multiday level because of the low reaction rate and accuracy.

The utilization of, smart card information to track student long haul travel exercises and examples, for example, the quantity of average day by day excursion chains, basic boarding and alighting time and excursion begin/end times, offers a significantly more advantageous and productive information source.

The goals of this study are to aid both travel offices and transportation analysts by:

1) Creating a novel information mining strategy to concentrate individual student travel examples and travel normality

2) Guaranteeing these information mining calculations are equipped for transforming gigantic savvy card datasets inside a fair slipped by time.

The rest of this paper is sorted out as takes after. The information utilized as a part of this study is initially presented.

DATA MINING IN HEALTH CARE

SEEMA DABADI (15MCA24)
JANUKACHHETRI (15MCA07)

Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. Data mining is the analysis of large data sets to discover patterns and use those patterns to forecast or predict the likelihood of future events.

The Three Systems Approach: The most effective strategy for taking data mining beyond the realm of academic research is the three systems approach. Implementing all three systems is the key to driving real-world improvement with any analytics initiative in healthcare.

The three systems are:

The analytics system. This system includes the technology and the expertise to gather data, make sense of it and standardize measurements

The best practice system. The best practice system involves standardizing knowledge work—systematically applying evidence-based best practices to care delivery.

The adoption system. This system involves driving change management through new organizational structures.

When these principles are in place, we have seen clients make some very energizing progress. Once they implement the analytics foundation to mine the data and they have the best practices and organizational systems in place to make data mining insights actionable, they are now ready to use predictive analytics in new and innovative ways.

Data Mining to Improve Primary Care Reporting

The first initiative mines historical EDW data to enable primary care providers (PCPs) to meet population health regulatory measures. This clinic's PCPs must demonstrate to regulatory bodies that they are giving the appropriate screenings and treatment to certain populations of patients. The EDW and analytics applications have enabled the PCPs to track their compliance rate and to take measures to ensure patients receive needed screenings. The Health Catalyst Advanced Application for Primary Care shows trending of compliance rates and specific measurements over time.

Data Mining to Predict Patient Population Risk

The second initiative involves applying predictive algorithms to EDW data to predict risk within certain populations. This process of stratifying patients into high-, medium- or low-risk groups is key to the success of any population health management initiative.

Interestingly, some patients carry so much risk that it would be cheaper to pre-emptively send a physician out to make a house call rather than waiting for that patient to come in for a crisis appointment or emergency room visit. The clinic needed to be able to identify these high-risk patients ahead of time and focus the appropriate resources on their care.

Data Mining to Prevent Hospital Readmissions

Reducing 30- and 90-day readmissions rates is another important issue health systems are tackling today. We have used data mining to create algorithms that identity those patients at risk for readmission.

When your health system has an adequate historical data set—i.e., you have adequate data about patients with certain conditions who were readmitted within 30 or 90 days—you can mine that data to

create an accurate predictive algorithm. The following is a high-level description of steps to learn from a historical cohort and create an algorithm.

Define a time period (the parameters of the historical data).

Identify all of the patients flagged for readmission in that time period.

Find everything those patients have in common (lab values, demographic characteristics, etc.).

Determine which of these variables has the most impact on readmissions. You can do this mathematically using a variety of statistical models.

Data mining can also help this health system streamline its efforts by evaluating the relative efficacy of each best practice. For example, if a case manager only has time to apply some of the interventions to a patient, which intervention or combination of interventions will have the most impact?

This brief case study is illustrative of what applying data mining in the real world is all about. If the health system had waited until its stars were perfectly aligned before getting started on its initiative, it might still be waiting today. Perhaps that's why data mining so often doesn't make it out of the academic lab and into everyday clinical practice. But this is the type of effort that is required—the determination to iterate step by step in a process of continuous quality improvement.

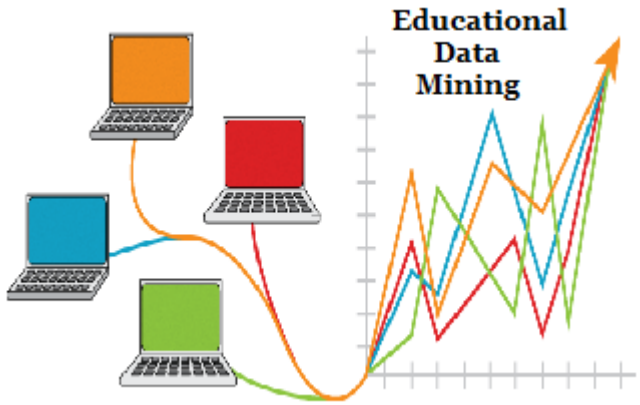
EDUCATIONAL DATA MINING

SANTOSINIDAS(15MCA23)

DESAI D.A(15MCA05)

"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."

Educational Data Mining (EDM) describes a research field concerned with the application of data mining, machine learning and statistics to information generated from educational settings (e.g., universities and intelligent tutoring systems).



Educational Data Mining focuses on developing new tools and algorithms for discovering data patterns.

EDM develops methods and applies techniques from statistics, machine learning, and data mining to analyze data collected during teaching and learning.

EDM tests learning theories and informs educational practice.

Educational data mining is emerging as a research area with a suite of computational and psychological methods and research approaches for understanding how students learn.

Data collected from online learning systems can be aggregated over large numbers of students and can contain many variables that data mining algorithms can explore for model building.

Phases of Educational Data Mining:

EDM generally consists of four phases:

1. The first phase of the EDM process (not counting pre-processing) is discovering relationships in data. This involves searching through a repository of data from an educational environment with the goal of finding consistent relationships

between variables. Several algorithms for identifying such relationships have been utilized, including classification, regression, clustering, factor analysis, social network analysis, association rule mining, and sequential pattern mining.

2. Discovered relationships must then be validated in order to avoid overfitting.
3. Validated relationships are applied to make predictions about future events in the learning environment.
4. Predictions are used to support decision-making processes and policy decisions.

Types of EDM methods:

- Prediction
  - Classification
  - Regression
  - Density estimation
- Clustering
- Relationship mining
  - Association rule mining
  - Correlation mining
  - Sequential pattern mining
  - Causal data mining
- Distillation of data for human judgment
- Discovery with models

Goals of EDM:

1. Predicting students' future learning behavior by creating student models that incorporate such detailed information as students' knowledge, motivation, metacognition, and attitudes;
2. Discovering or improving domain models that characterize the content to be learned and optimal instructional sequences;
3. Studying the effects of different kinds of pedagogical support that can be provided by learning software; and
4. Advancing scientific knowledge about learning and learners through building computational models that incorporate models of the student, the domain, and the software's pedagogy.

AUTOMATED PERSONALITY CLASSIFICATION USING DATA MINING TECHNIQUES

NETHRAVATHI.M(15MCA17)
CHITHRA.M.S(15MCA04)

Personality identification of a human being by their nature an old technique. Earlier these were done manually by spending lot of time to predict the nature of the person. Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. Methods used to analyze the data include surveys, interviews, questionnaires, classroom activities, shopping website data, social network data about the user experiences and problems they are facing. But these traditional methods are time consuming and very limited in scale. The manual analysis does not make sense of analyzing user learning experiences which are huge in volume with different Internet slang and the timing of the user posting on the web. The sentiment analysis of the user collected data does not cover much relevant experience because even for a human judge to determine what user problems a data indicates is a more complicated task than to determine just the sentiment of a data. Humans are prone to biases and prejudices which may affect the accuracy of their judgments. Also, certain features of a Facebook profile or other social networks text data are difficult for humans to grasp. For example, while the number of Facebook friends is clearly displayed on the profile, it is more difficult for a human to determine features such as the network density.

Approaches To Automated Personality Classification Using Data Mining Techniques:

1. Novel approaches to automated personality classification: Ideas and their potentials:

   This work proposes several new research directions regarding the problem of Automated Personality Classification (APC). Firstly, possible improvements of the existing solutions to the problem of APC, for which different combinations of the APC corpora, psychological trait measurements and learning algorithms are used is investigated. Afterwards, the extensions of the APC problem and the related tasks, such as dynamical APC and detecting personality inconsistency in a text is considered. This entire research was performed in the context of social networks and the related data mining mechanisms.

2. Educational Game (Detecting personality of players in an educational game):

   The educational Data Mining is to develop the methods for student modeling based on educational data, such as; chat conversation, class discussion, etc. On the other hand, individual behavior and personality play a major role in Intelligent Tutoring Systems (ITS) and Educational Data Mining (EDM). Thus, to develop a user adaptable system, the student's behaviors that occurring during interaction has huge impact on EDM and ITS. In this work, a novel data mining techniques and natural language processing approaches for automated detection student's personality and behaviors in an educational game (Land Science) is introduced where students act as interns in an urban planning firm and discuss in groups their ideas.

3. A System for Personality and Happiness Detection:

This work proposes a platform for estimating personality and happiness. Starting from Eysenck's theory about human's personality, authors seek to provide a platform for collecting text messages from social media (Whatsapp), and classifying them into different personality categories. Although there is not a clear link between personality features and happiness, some correlations between them could be found in the future. In this work, the platform developed is described, and as a proof of concept, different sources of messages is used to see if common machine learning algorithms can be used for classifying different personality features and happiness.

4. Using Twitter Content to Predict Psychopathy:

An ever-growing number of users share their thoughts and experiences using the Twitter micro logging service. Although sometimes dismissed as containing too little content to convey significant information, these messages can be combined to build a larger picture of the user posting them. One particularly notable personality trait which can be discovered this way is psychopathy: the tendency for disregarding others and the rule of society. In this approach, techniques are explored to apply data mining towards the goal of identifying those who score in the top 1.4% of a well-known psychopathy metric using information available from their Twitter accounts.

DATA WAREHOUSING FOR ROUGH WEB CACHING AND PRE-FETCHING

MARY RESHMA JENNIFER J (15MCA15)

NIKITHA J (15MCA18)

Web caching and prefetching are the most popular techniques that play a key role in improving the Web performance by keeping web objects that are likely to be visited in the near future closer to the client. Web caching can work independently or integrated with the web prefetching. The Web caching and prefetching can complement each other since the web caching exploits the temporal locality for predicting revisiting requested objects, while the web prefetching utilizes the spatial locality for predicting next related web objects of the requested Web objects.

Caching and pre-fetching is middle-aged technology widely used in many areas such as Database Systems and Operating Systems. Presently, the World Wide Web becomes another attractive area in applying caching and pre-fetching. The hit rate and byte hit rate are two extensively applied performance metrics in Web caching [1, 2]. Hit rate is the ratio of the number of requests that reach the proxy cache and the total number of requests. Byte hit rate is the ratio of the number of bytes that reach the proxy cache and the total number of bytes requested. As Yang and Zhang has indicated: "Fractional network traffic is also defined to measure the increased network load".

Fractional latency is the ratio between the observed latency with and without a caching or prefetching system. Fractional network traffic is the ratio between the number of bytes that are transmitted from Web servers to the proxy and the total number of bytes requested. Clearly, the lower the fractional latency and network traffic, the better the performance. For example, the fractional latency of 30% achieved by a caching system means the caching system saves 70% of the latency.Web caching and Web pre-fetching are two essential techniques used to reduce the visible response time identified by users.

Web Caching: Web caching is one of the most successful solutions for improving the performance of Web-based system. In Web caching, the popular web objects that likely to be visited in the near future are stored in positions closer to the user like client machine or proxy server. Thus, the web caching helps in reducing Web service bottleneck, alleviating of traffic over the Internet and improving scalability of the Web system. The Web caching has three attractive advantages to web participants, including end users, network managers, and content creators: i) The Web caching decreases user perceived latency. ii) The Web caching reduces network bandwidth usage. iii) The Web caching reduces loads on the origin servers.

Web Prefetching: Web prefetching is another very effective technique, which is utilized to complement the Web caching mechanism. The web prefetching predicts the web object expected to be requested in the near future, but these objects are not yet requested by users. Then, the predicted objects are fetched from the origin server and stored in a cache. Thus, the web prefetching helps in increasing the cache hits and reducing the user-perceived latency.

An implementation of a data warehouse for rough Web caching and pre-fetching, which not only considers the caching effect in the Web environment, but also evaluates the pre-fetching rules. Explicitly, a normalized profit function is formulated to evaluate the profit from caching an object either no-cache or cache object according to some pre-fetching rule. Teng et al used the normalized profit function devised an innovative Web cache replacement algorithm, so called as Algorithm IWCP (standing for the Integration of Web Caching and Pre-fetching). They evaluate the performance of Algorithm IWCP under several circumstances by using an event-driven simulation. A proxy server is usually located at the edge of a LAN, intercepting HTTP requests and responses between clients and Web servers.If it found that the requested object is already stored in its cache, returns the object to the user. Otherwise, it goes to the original server on behalf of the user, grabs the object, stores it in its cache, and returns the object to the user. An advantage of Web proxy caching and pre-fetching is that all clients within the Local Area Network (LAN) can share objects stored in the cache.

In order to calculate the benefits of pre-fetching, the common currency that will be used to measure the benefit must be selected. Since the ultimate goal of this study is to mask document latency, latency is selected as the currency and to attempt to minimize it.

Conclusion: The results presented have managed to determine the possible Web pre-fetching according to hit ratio and hit links. Furthermore, it narrows down and increases the performances on overall links, protocol and size byte rather than just links. In order to incorporate pre-fetching into a Web caching system, proxy server should have the pre-fetch engine to narrow down the possible links at the local level. The effectiveness of this pre-fetching and caching algorithm depends and may vary on the usage and the type of search that is available. For example, Google has implemented the most frequent user search. By the time user finishes typing the word they are looking for, a list of all the hits is listed.

OPEN SOURCE DATA MINING TOOLS

SUREKHA (15MCA28)
RANJITHA (15MCA21)

Data mining, also known as knowledge discovery from databases, is a process of mining and analyzing enormous amounts of data and extracting information from it. Best Open Source Data Mining Tools

Rapid Miner (formerly known as YALE)

Written in the Java Programming language, this tool offers advanced analytics through template-based frameworks. A bonus: Users hardly have to write any code. Offered as a service, rather than a piece of local software, this tool holds top position on the list of data mining tools. In addition to data mining, Rapid Miner also provides functionality like data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment. What makes it even more powerful is that it provides learning schemes, models and algorithms from WEKA and R scripts.

WEKA

The original non-Java version of WEKA primarily was developed for analyzing data from the agricultural domain. With the Java-based version, the tool is very sophisticated and used in many different applications including visualization and algorithms for data analysis and predictive modeling. Its free under the GNU General Public License, which is a big plus compared to Rapid Miner, because users can customize it however they please.WEKA supports several standard data mining tasks, including data preprocessing, clustering, classification, regression, visualization and feature selection. WEKA would be more powerful with the addition of sequence modeling, which currently is not included.

R-Programming

What if I tell you that Project R, a GNU project, is written in R itself? It's primarily written in C and FORTRAN. And a lot of its modules are written in R itself. It's a free software programming language and software environment for statistical computing and graphics. The R language is widely used among data miners for developing statistical software and data analysis. Ease of use and extensibility has raised R's popularity substantially in recent years. Besides data mining it provides statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others.

Orange

Python is picking up in popularity because it's simple and easy to learn yet powerful. Hence, when it comes to looking for a tool for your work and you are a Python developer, look no further than Orange, a Python-based, powerful and open source tool for both novices and

experts. You will fall in love with this tool's visual programming and Python scripting. It also has components for machine learning, add-ons for bioinformatics and text mining. It's packed with features for data analytics.

KNIME

Data preprocessing has three main components: extraction, transformation and loading. KNIME does all three. It gives you a graphical user interface to allow for the assembly of nodes for data processing. It is an open source data analytics, reporting and integration platform. KNIME also integrates various components for machine learning and data mining through its modular data pipelining concept and has caught the eye of business intelligence and financial data analysis.
Written in Java and based on Eclipse, KNIME is easy to extend and to add plugins. Additional functionalities can be added on the go. Plenty of data integration modules are already included in the core version

Apache Mahout

Mahout is primarily a library of machine learning algorithms that can help in clustering, classification and frequent pattern mining. It can be used in a distributed mode that helps easy integration with Hadoop. Mahout is currently being used by some of the giants in the tech industry like Adobe, AOL, Drupal and Twitter, and it has also made an impact in research and academics. It can be a great choice for anyone looking for easy integration with Hadoop and to mine huge volumes of data.

SPATIAL DATA MINING

KAVYASURESH(15MCA08)

ROOPA M S(15MCA22)

Introduction

Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases. The complexity of spatial data and intrinsic spatial relationships limits the usefulness of conventional data mining techniques for extracting spatial patterns. Efficient tools for extracting information from geospatial data are crucial to organizations which make decisions based on large spatial datasets, including NASA, the National Imagery and Mapping Agency (NIMA), the National Cancer Institute (NCI), and the United States Department of Transportation (USDOT). These organizations are spread across many application domains including ecology and environmental management, public safety, transportation, Earth science, epidemiology, and climatology

Spatial Data Mining Tasks

- ✓ **Classification**– finds a set of rules which determine the class of the classified object according to its attributes
- ✓ **Association rules** – find (spatially related) rules from the database. Association rules describe patterns, which are often in the database.
- ✓ **Discriminate rules** – Describe differences between two parts of database e. g. find differences between cities with high and low unemployment rate.
- ✓ **Clustering** – Groups the object from database into clusters in such a way that object in one cluster are similar and objects from different clusters are dissimilar.
- ✓ **Trend detection** – Finds trends in database. A trend is a temporal pattern in some time series data. A spatial trend is defined as a pattern of change of a non-spatial attribute in the neighborhood of a spatial object.

Spatial Data Mining Techniques

There is no unique way of classifying SDM techniques. Various kinds of patterns can be discovered from databases and can be presented in different forms. Based on general data mining, tasks can be classified into two main categories: descriptive data mining and predictive data mining.This paper focuses on the organization of the particular spatial data mining techniques as .

- ✓ **Clustering and Outlier Detection** .

Spatial clustering is a process of grouping a set of spatial objects into groups called clusters. Objects within a cluster show a high degree of similarity, whereas the clusters are as much dissimilar as possible. Clustering is a very well known technique in statistics and clustering algorithm to deal with the large geographical datasets. Clustering algorithms can be separated into four general categories: partitioning method, hierarchical method, densitybased method and grid-based method. The categorization is based on different cluster definition techniques.

- ✓ **Association and Co-Location**

When performing clustering methods on the data, we can find only characteristic rules, describing spatial objects according to their non-spatial attributes. In many situations we want to discover spatial rules that associate one or more spatial objects with others. However, one of the biggest research challenges in mining association rules is the development of methods for selecting potentially interesting rules from among the mass of all discovered rules.

- ✓ **Classification**

Every data object stored in a database is characterized by its attributes. Classification is a technique, which aim is to find rules that describe the partition of the database into an explicitly given set of classes. Classification is considered as predictive spatial data mining, because we first create a model according to which the whole dataset is analyzed.

- ✓ **Trend-Detection**

A spatial trend is a regular change of one or more non-spatial attributes when spatially moving away from a start object. Therefore, spatial trend detection is a technique for finding patterns of the attribute changes with respect to the neighborhood of some spatial object.

Spatial Data Mining Applications

- ✓ **Spatial Trend Detections in GIS**

Spatial trends describe a regular change of non-spatial attributes when moving away from certain start objects. Global and local trends can be distinguished. To detect and explain such spatial trends, e.g. with respect to the economic power, is an important issue in economic geography.

- ✓ **Spatial Characterization of Interesting Regions**

Another important task of economic geography is to characterize certain target regions such as areas with a high percentage of retirees. Spatial characterization does not only consider the attributes of the target regions but also neighboring regions and their properties.

Conclusion

This paper presented the techniques of spatial data mining in the following four categories Clustering and Outlier Detection, Association and Co-Location, Classification and Trend-Detection. It also discussed some trends and applications of spatial data mining. Finally, it identified research needs for spatial data mining.