On Line Analytical Mining of Web Usage Data Warehouse

Ratnesh Kumar Jain[1], Dr. Suresh Jain[2], and Dr. R. S. Kasana[1]

[1]Department of Computer Science & Applications,
Dr. H. S. Gour, University, Sagar, MP (India)
jratnesh@rediffmail.com
[2] Department of Computer Engineering, Institute of Engineering & Technology,
Devi Ahilya University, Indore, MP (India)
suresh.jain@rediffmail.com

Abstract

With the rapid growth of World Wide Web and its access, web logs are emerged as huge data repository of page access information which when mined properly can help improve system design, better marketing decisions etc. On Line Analytical Mining (OLAM) is utilized in many areas. But it is not utilized in the field of web usage mining. If data in web logs can be represented in the form of data warehouse (data cube) we can utilize OLAM. The primary requirement in the construction of Multidimensional Data Cube is identification of dimensions and measures. In this paper we present the low cost technique of Web log data mining by multi-dimensional analysis of Web log data. At first, we suggested the dimensions and measures in web usage data warehouse and then we proposed how does On Line Analytical Mining (OLAM) can be applied on web usage data warehouse? Which type of pattern analysis it can perform on web logs? And what are the advantages and disadvantages of using OLAM on web usage data warehouse.

Keywords: Web usage mining, OLAM, Web usage Data warehouse, data cube, dimension and measure.

1. Introduction

The huge development in the technology provides a great boost to the database and information industry, and makes a large number of databases and information repositories available for transaction management, information retrieval, and data analysis. Data can now be stored in many different kinds of databases and information repositories. One such data repository architecture that has emerged is the data warehouse.

We can define Data warehouse as "a huge repository of multiple heterogeneous data sources organized under a unified schema at a single site in order to facilitate management decision making" [1]. Once the Data warehouse is constructed we apply intelligent methods called data mining techniques to extract data patterns. In general there are two levels of data mining-

- Descriptive level
- Predictive level

Descriptive level is more interactive, temporary and query driven [1]. In descriptive level data mining data are presented in multidimensional data cubes and traditional analysis and reporting tools that are provided by OLAP techniques are used [2]. It is known as OLAM (On Line Analytical Mining). Predictive level data mining is however more automatic. In predictive level different data mining algorithms are used to discover new implicit patterns.

The World Wide Web provide rich, worldwide, on-line information services, where data objects are linked together to facilitate interactive access. Users navigate from one web page to another to search the information of his or her interest. The users' accesses to the web pages are recorded into a file called web logs. Web logs provide a huge repository of page access information which when mined properly can help improve system design, better marketing decisions etc. If data in web logs can be represented in the form of data warehouse (data cube) we can apply the OLAP techniques to analyze web usage patterns. Hence data warehouses, data cubes and OLAP techniques have emerged as a scalable and flexible approach to mining Web log files. Han et al. [3] has shown that some of the analysis needs of Web usage data can be done using data warehouses and OLAP techniques.

2. Problem Statement

Most of the algorithm used for web usage mining reveal only frequency count or frequent access sequences. Some of these algorithms have limitations with regard to the size of the web log files, whether it is physical in size or practical in time because of the low speed of the analysis. In order to reduce the size these algorithms make assumptions that decrease the accuracy of mining. Also, the contents of most web sites change over time, making some parts of Web logs irrelevant for the current analysis. In addition, the goal of analysis may change over time with the business needs. Hence we can say that current web log analysis tools are limited in their performance, the comprehensiveness and depth of their analysis, scalability and the validity and reliability of their results. Therefore, on the one hand, there is a need, both in terms of memory and disk space, for scalability improvements, and, on the other hand, for the introduction of constraints (such as time constraints, concept hierarchies and frequent pattern templates) into mining algorithms, for discovering relevant and correct knowledge.

3. Data Warehousing and OLAM

Data warehouse generalizes and consolidate data in multidimensional space and provide on-line analytical processing (OLAP) tools for the interactive analysis which facilitates effective data generalization and data mining (OLAM) [4,5,6].

3.1 Data Warehousing

Data warehouse systems provide architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions. Data warehouse is constructed using following steps:

- Data Cleaning: Applied to remove noise and correct inconsistencies in the data.
- Data Integration: Merges data from multiple sources into a coherent data store that is data warehouse.
- Data Transformation: Includes normalization and aggregation. Contributes toward the success of the mining process.
- Data Loading: Process that move data from one data system to another system, especially, make the data become accessible for the data warehouse.

- Data Transformation: Includes normalization and aggregation. Contributes toward the success of the mining process.
- Data Loading: Process that move data from one data system to another system, especially, make the data become accessible for the data warehouse.

- Periodic Data Refreshing: A data warehouse is populated by performing an initial loading. Then, it is regularly updated during the periodic execution of a refreshment process.

The data in the data warehouse are subject-oriented, nonvolatile, historical and stored in the summarized form [7]. Generally data warehouse is modeled by a multidimensional database structure called **data cubes** (see figure 1 a, b and c)**.** A data cube is defined by dimensions and facts. **Dimensions** can be defined as perspectives (major subjects) or entities with respect to which an organization wants to keep records such as time, locality, item, supplier etc. Each dimension corresponds to an attribute or a set of attributes in the schema, and each **cell** stores the value of some aggregate **measure** called **fact.** A data cube measure is a numerical function that can be evaluated at each point in the data cube space. A measure value is computed for a given point by aggregating the data corresponding to the respective dimension-value pairs defining the given point.

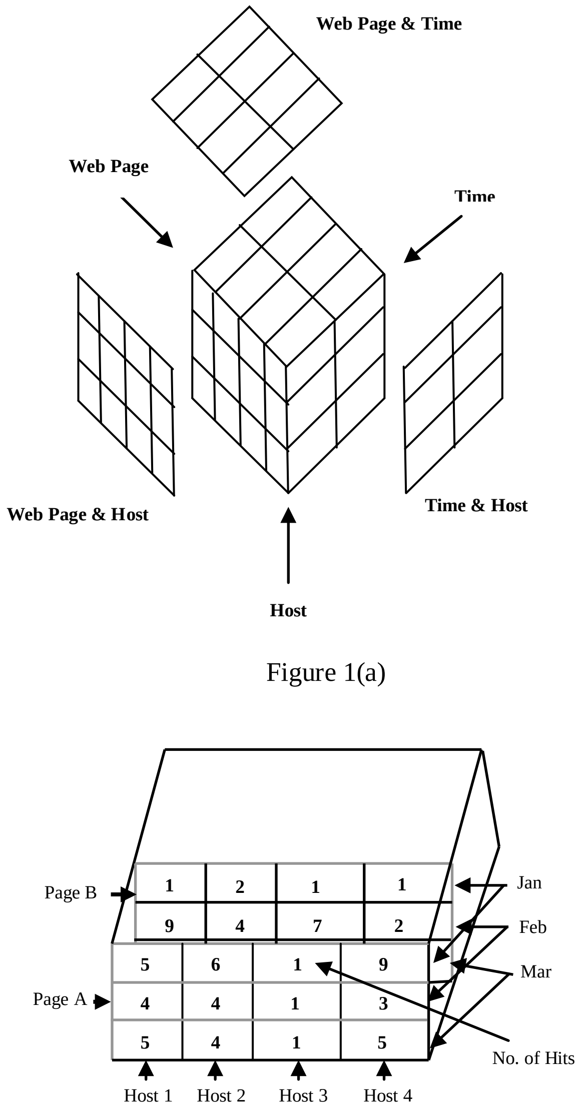


Figure 1(a)

Figure 1(b)

Figure 1, Example data cubes

3.2 On-Line Analytical Mining (OLAM)

Online Analytical Mining also called OLAP mining integrates on-line analytical processing (OLAP) with data mining and mining knowledge in multidimensional databases. Among the many different approaches of data mining, OLAM is particularly important for the following reasons:

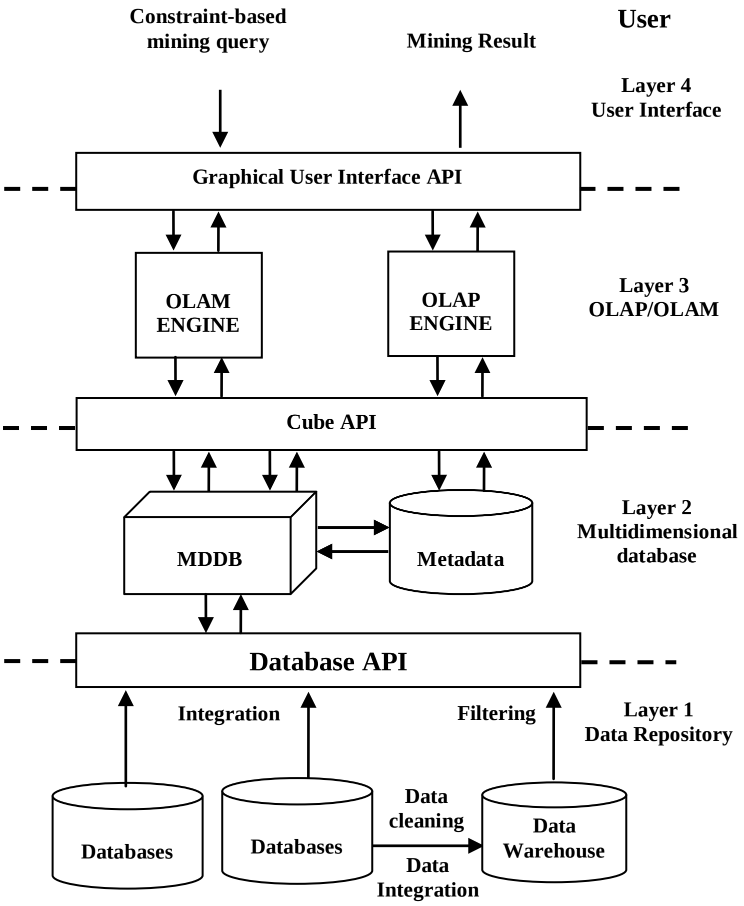


Figure 2, A combined Architecture of OLAM and OLAP**.**

- High quality of data in Data Warehouses.
- Availability of many data warehouses based information processing techniques.
- OLAPbased exploratory data analysis.
- Facility to integrate OLAP with multiple data mining functions on-line.

The architecture for On-Line Analytical Mining is shown in figure 2. Online Analytical Processing or OLAP is part of the broader category business intelligence. OLAP is an approach to quickly provide answers to analytical queries that are based on multi-dimensional data model. Since data warehouses provide multidimensional data views and the precomputation of summarized data in the form of data cubes. As shown in the figure OLAM architecture is similar to OLAP architecture. Since an OLAM server may perform multiple data mining tasks, such as concept description, association, classification, prediction, clustering, time-series analysis, and so on, it usually consists of multiple data mining modules and is more sophisticated than an OLAP server.

OLAP has emerged as a powerful paradigm for strategic analysis of data warehouse systems. The typical applications of OLAP are in business reporting for sales, marketing, management reporting, business process management (BPM), budgeting and forecasting, financial reporting and similar areas. The term OLAP was created as a slight modification of the traditional database term OLTP (**Online Transaction Processing**). OLAP includes Summarization, Consolidation, Aggregation, and facility to view information from different angles[10,11].

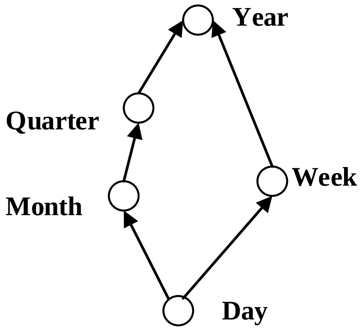


Figure 3, Concept hierarchies for time dimension.

In the multidimensional data cube model each dimension contains multiple levels of abstraction (granularity) called concept. OLAP uses a concept hierarchy (figure 3 shows concept hierarchy for time dimension) that defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts. The operations that can be performed in OLAP use this concept hierarchy to provide users with the flexibility to view data from different perspectives and allowing interactive querying and analysis of the data at hand. And by this way it provides a user-friendly environment for interactive data analysis. The OLAP operations that help in interactive data analysis are:

3.2.1 Slice

The slice operation is based on selecting one dimension and focusing on a portion of a cube.

3.2.2 Dice

The dice operation creates a sub-cube by focusing on two or more dimensions.

3.2.3 Roll-up

Roll-up, also called aggregation or dimension reduction, allows the user to move to the higher aggregation level.

3.2.4 Drill-down

The drill-down operation is the reverse of a roll-up and represents the situation when the user moves down the hierarchy of aggregation, applying a more detailed grouping.

3.2.5 Pivoting

Pivoting, or rotation, changes the perspective in presenting the data to the user.

To analyze large and detailed databases, users have two options: to purchase a specialized OLAP product or to use OLAP cube functionality available in other commonly used products such as spreadsheets or statistical software packages. Selecting the most suitable OLAP product is a very complex process. Since these products are rather very expensive, it has to be done using the more formal implementation process. The selection of an optimal OLAP product should follow the user needs analysis, system selection, cost/benefit analysis and the whole product life-cycle methodology elaborated for other IT products. However, there is a low cost alternative. Excel users can create OLAP cubes using pivot tables and SPSS users can generate OLAP cubes with SPSS reports. Both Excel and SPSS users can easily create OLAP cubes and manipulate them in a very efficient way. In addition, each of these products offers some specific functionality. For example, in Excel, users can create graphs corresponding to OLAP cubes, whereas SPSS provides more statistics for analyzing data in OLAP cubes.

4. OLAM on Web Usage Data

In this section we present how web log data can be converted into data warehouse (some of the researchers named it webhouse [8, 9]) and how can we apply OLAM technique to mine interesting patterns?

4.1 Web Usage Data warehouse and OLAM

Data stored in web logs can be cleaned, preprocessed and integrated into one huge data repository that is data warehouse. And then OLAP techniques can be applied to mine useful access patterns. The whole process is shown in figure 4.

4.1.1 Data Preprocessing:

Data collected in the web logs are filtered to remove irrelevant information and a relational database is created containing the meaningful remaining data. This database facilitates information extraction and data summarization based on individual attributes like user, locality, time etc. Some of the filtering that is done by many web log mining tools but not adopted in OLAM is:

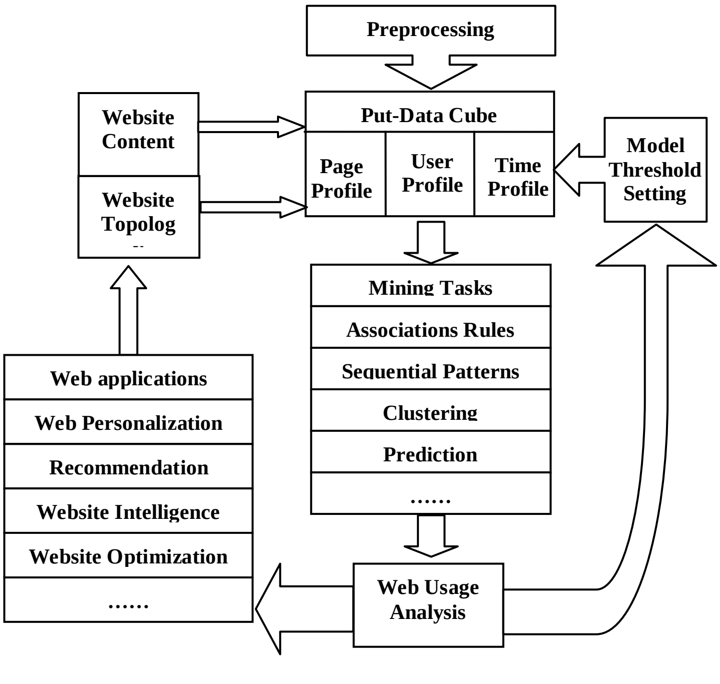


Figure 4, The multidimensional model for web usage mining.

- Remove the logs about the graphics, sound and video pages. The logic behind this is that these pages are part of a web page so they are requested in order to display the actual page. But in OLAM we are interested to keep these entries because we believe that entry in the log about these pages can give us interesting clues regarding web site structure, traffic performance, as well as user motivation. Also, one user action can generate multiple and some of them are requests for media pages. Some of these logs are important to realize the intended action of the user.
- Elimination of log entries generated by web agents like web spiders, indexers, link checkers, or other intelligent agents that pre-fetch pages for caching purposes. But we are interested in keeping these logs as they are helpful to analyze web agents' behavior on a site and compare the traffic generated by these automated agents with the rest of the traffic.

- Our concentration in this step is to reduce the elimination because most of the data are relevant in any way. Also, in data warehouse we do not have the problem of space that is the biggest reason of the data elimination in these mining algorithms. The data filtering we adopted mainly transforms the data into a more meaningful representation.

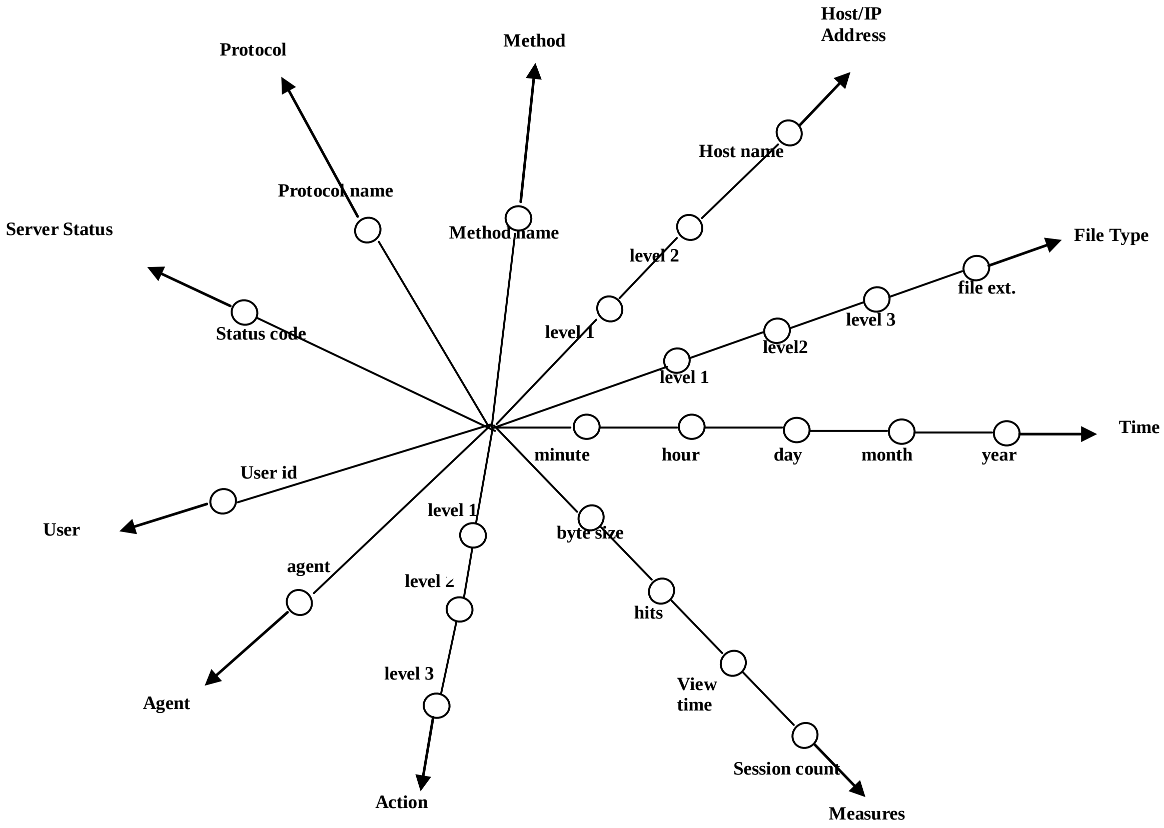
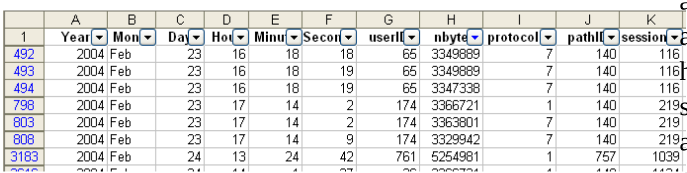


Figure 5, Fragment of Relational Database containing preprocessed web logs



	A	B	C	D	E	F	G	H	I	J	K
1	Year	Mon	Day	Hou	Minu	Secon	userID	nbyte	protocol	pathID	session
492	2004	Feb	23	16	18	18	65	3349889	7	140	116
493	2004	Feb	23	16	18	19	65	3349889	7	140	116
494	2004	Feb	23	16	18	19	65	3347338	7	140	116
798	2004	Feb	23	17	14	2	174	3366721	1	140	219
803	2004	Feb	23	17	14	2	174	3363801	7	140	219
808	2004	Feb	23	17	14	9	174	3329942	7	140	219
3183	2004	Feb	24	13	24	42	761	5254981	1	757	1039

Figure 6, Example concept hierarchy for Web log data cube

After the cleaning and transformation of the web log entries, the web log is loaded into a relational database and some additional data, such as time spent by event, is calculated. Figure 5, shows relational database that contain preprocessed web logs in the form of rows and columns.

4.1.2 Construction of Data Cube

As the preprocessing task is finished we construct a multidimensional data cube and load the preprocessed data from relational data base into it. To construct a web log data cube we must have the knowledge of all the dimensions in which web log data can be arranged and all the possible attributes of each dimension that describe facts. For example the time dimension may have the attributes such as minute, hour, day, month, and year. Attributes of a dimension may be related by partial order indicating a hierarchical relationship among the dimension attributes. Each dimension is defined on a concept hierarchy to facilitate generalization and specialization along the dimension. Figure 6 shows all the dimensions, attributes and concept hierarchy of each dimension.

Each line in the figure 6 represents a dimension and each circle in the line represents an attribute. Each line is representing the concept hierarchy for that dimension. This starnet model in figure 6 shows 9 dimensions and four measures that can be constructed directly from Web log. For more business oriented analysis, other attributes from other data sources, such as user profile and referrer logs, can be added to this model. Additional attributes may include user demographic data, such as address, age, education, and income and referrer data such as referrer address, local URL can also be added.

After the construction of multi-dimensional data cube various OLAP techniques are applied to provide further insight of any target data set from different perspectives and at different conceptual levels. This counts as summarization function in data mining as well. Some typical summarization includes the following: Byte transferred, Hits count, View time, Session count, Domain summary, Event summary, Request summary.

The multi-dimensional structure of the data cube provides great flexibility to manipulate the data and view it from different perspectives. The sum cells allow quick summarization at different levels of the concept hierarchies defined on the dimension attributes.

Building this web log data cube allows the application of OLAP (On-Line Analytical Processing) operations, such as drill-down, roll-up, slice and dice, to view and analyze the web log data from different angles, derive ratios and compute measures across many dimensions.

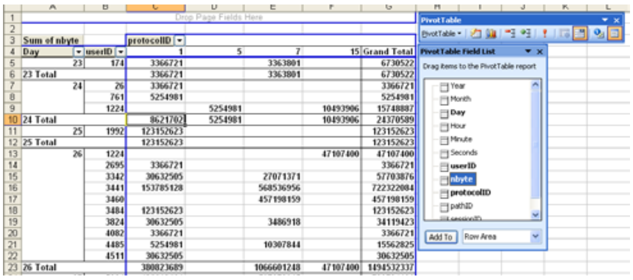
4.1.3 OLAM the OLAP mining on web log data

"OLAP is a method for interactively exploring multidimensional data." In multidimensional OLAP analysis, standard statistical measures (such as counts, sum of the bytes transferred) are applied to assist the user at each step to explore the interesting parts of the cube. Web log data cubes are constructed to give the user the flexibility of viewing data from different perspectives and performing ad hoc analytical quires. A user can use the data cube to analyze how overall usage of Web site has changed in the last quarter, to check whether most server requests have been answered, hopefully with expected or low level of errors. If some weeks or days are worse than the others, the user might navigate further down into those levels, always looking for some reason to explain the observed anomalies. At each step, the user might add or remove some dimension, changing their perspective, select subset of the data at hand, drill down, or roll up, and then inspect the new view of the data cube again. Each step of this process signifies a query or hypothesis, and each query follows the result of the previous step[10,11]. The knowledge that can be discovered is represented in the form of rules, tables, charts, graphs, and other visual presentation forms.

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. Data mining functionalities, and the kinds of patterns they can discover, are described below:

- **Class Description**:

Data can be associated with classes or concepts. Such as classes of agent include Mozilla, NetScape Navigator etc. It is very helpful to describe individual classes and concepts in summarized, concise and precise terms. Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived via



Sum of nbyte		protocolID				
Day	userID	1	5	7	15	Grand Total
23	174	3366721		3363801		6730522
23 Total		3366721		3363801		6730522
24	26	3366721				3366721
	761	5254981				5254981
	1224		5254981		10493906	15748887
24 Total		8621702	5254981		10493906	24370589
25	1992	123152623				123152623
25 Total		123152623				123152623
26	1224				47107400	47107400
	2695	3366721				3366721
	3342	30632505		27071371		57703876
	3441	153785128		568536956		722322084
	3460			457198159		457198159
	3484	123152623				123152623
	3824	30632505		3486918		34119423
	4082	3366721				3366721
	4485	5254981		10307844		15562825
	4511	30632505				30632505
26 Total		380823689		1066601248	47107400	1494532337

Figure 7, Example data cube created using Pivot Table having Day, UserID, ProtocollID as Dimensions and Number of bytes transferred as measure.

- **Data Characterization**:

It consists of finding rules that summarize general characteristics of a set of user-defined data. The rules are generated from a generalized data cube produced using the web log data cube and the OLAP operations. For example, the traffic on a web server for a given type of media in a particular time of day can be summarized by a characteristic rule.

- **Data discrimination**:

It is done by comparison of the target class with one or a set of comparative classes (often called the contrasting classes). Comparison plays the role of examining the Web log data to discover discriminate rules, which summarize the features that distinguish the data in the target class from that in the contrasting classes. For example, to compare requests from two different web browsers, a discriminate rule summarizes the features that discriminate one agent from the other, like time, file type, etc.

- **Association:**

Mining frequent patterns leads to the discovery of interesting associations and correlations within data. Rule support and confidence are two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules. This function mines association rules at multiple-levels of abstraction. For example, one may discover the patterns that accesses to different resources consistently occurring together, or accesses from a particular place occurring at regular times. Typically, association rules are discarded as uninteresting if they do not satisfy both a minimum support threshold and a minimum confidence threshold.

- **Classification:**

Classification consists of finding a model that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data that is data objects whose class label is known. In web usage mining, classification consists of building a model for each given class based upon features in the web log data and generating classification rules from such models. The models are constructed by analyzing a training web log data set whose class label is known. The classification rules can be used to develop a better understanding of each class in the web log database, and perhaps restructure a web site or customize answers to requests(i.e. quality of service) based on classes of requests.

- **Prediction:**

Whereas classification predicts categorical (discrete, unordered) labels, Prediction models continuous-valued functions. That is, it is used to predict missing or unavailable numerical data values rather than class labels. Prediction involves predicting values or value distributions of an attribute of interest based on its relevance to other attributes. Both relevance analysis and predictive model construction need statistical analysis techniques. For example, the access to a new resource on a given day can be predicted based on accesses to similar old resources on similar days, or the traffic for a given page can be predicted based on the distribution of traffic on other pages in the server directory.

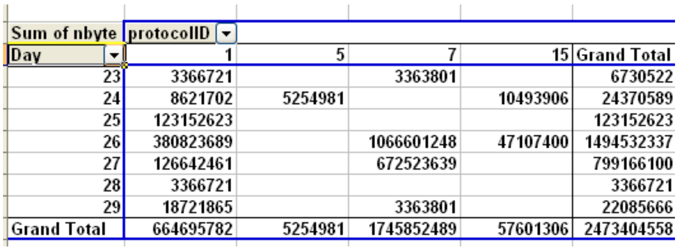
- **Evolution Analysis**:

It describes and models regularities or trends for objects whose behavior changes over time. Time-series analysis is to analyze data collected along time sequences to discover time-related interesting patterns, characteristics, trends, similarities, differences, periodicity, and so on. It may include data characterization, data discrimination, association, classification, and prediction. For example, time-series analysis of the web log data may disclose the patterns and trends of web page accesses in the last year and suggest the improvement of services of the web server.

The kind of patterns listed above can be find out by applying different OLAP operations such as slice, dice, drill-down, roll-up and pivoting. A GUI can be developed to automate the task of applying operations on cube and giving the results. We tried to explain the web log data cube analysis with the help of some examples in following section 4.2.

4.2 Illustrative Example

To illustrate the OLAP mining we used small set of web access logs downloaded free from the internet. This Web logs is of 23 feb, 2004 to 29 feb, 2004 having: 61 KB in size, 36,878 entries which includes 7851 unique host names, 2793 unique URLs and 32 different file types (extensions). Based on the file extensions these files can be grouped into audio, compressed, document, dynamic, html, images, java and video category. The IP addresses are grouped into domains. We considered nine dimensions and four measures to construct Web log OLAP data cubes. The dimensions are time, file, file type, host, method, protocol, agent, action and server status dimensions and measures are bytes and hits, view time and session count.



Sum of nbyte	protocolID				
Day	1	5	7	15	Grand Total
23	3366721		3363801		6730522
24	8621702	5254981		10493906	24370589
25	123152623				123152623
26	380823689		1066601248	47107400	1494532337
27	126642461		672523639		799166100
28	3366721				3366721
29	18721865		3363801		22085666
Grand Total	664695782	5254981	1745852489	57601306	2473404558

Figure 8, Dicing Data cube shown in figure 7 contains two dimensions day and protocolID.

Here we illustrate a simple yet typical example of summarization using OLAP technique (using MS Excel) on web usage data warehouse. In this example, we are interested in finding out how many bytes were transferred (**Web traffic analysis**) on each day of the month over a whole year for a particular protocol, user wise.

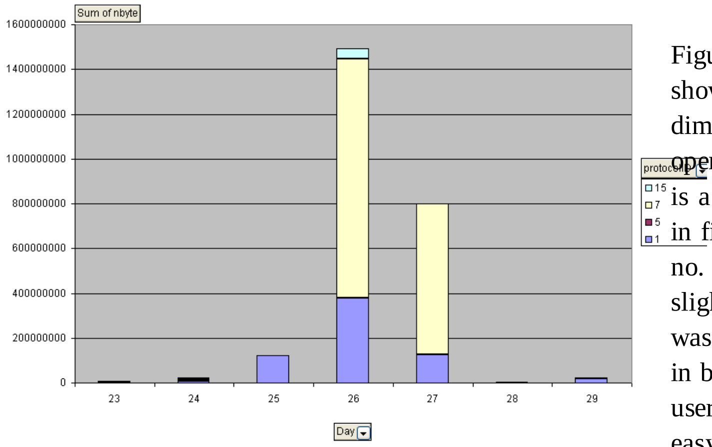
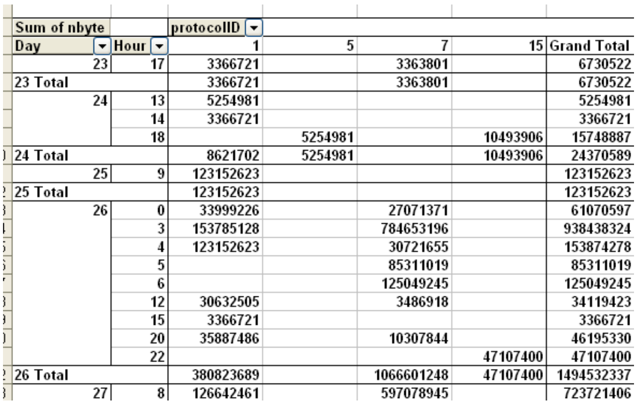
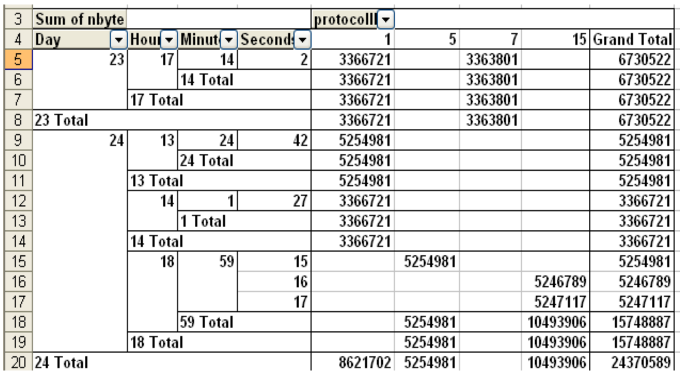


Figure 9, Graphical representation of the data cube shown in figure 8.



Sum of nbyte		protocolID				
Day	Hour	1	5	7	15	Grand Total
23	17	3366721		3363801		6730522
23 Total		3366721		3363801		6730522
24	13	5254981				5254981
	14	3366721				3366721
	18		5254981		10493906	15748887
24 Total		8621702	5254981		10493906	24370589
25	9	123152623				123152623
25 Total		123152623				123152623
26	0	33999226		27071371		61070597
	3	153785128		784653196		938438324
	4	123152623		30721655		153874278
	5			85311019		85311019
	6			125049245		125049245
	12	30632505		3486918		34119423
	15	3366721				3366721
	20	35887486		10307844		46195330
	22				47107400	47107400
26 Total		380823689		1066601248	47107400	1494532337
27	8	126642461		597078945		723721406

Figure 10, Resultant Data Cube after Drill down to Hours, Minutes and seconds in the time dimension in the data cube given in figure 8.

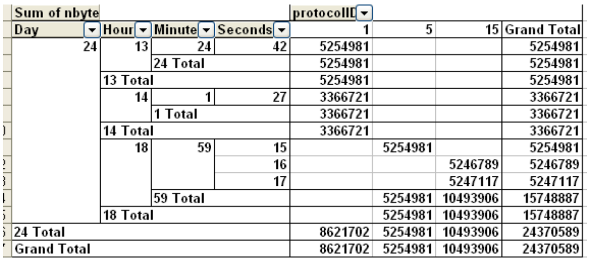


3	Sum of nbyte				protocolI				
4	Day	Hour	Minut	Second	1	5	7	15	Grand Total
5	23	17	14	2	3366721		3363801		6730522
6			14 Total		3366721		3363801		6730522
7		17 Total			3366721		3363801		6730522
8	23 Total				3366721		3363801		6730522
9	24	13	24	42	5254981				5254981
10			24 Total		5254981				5254981
11		13 Total			5254981				5254981
12		14	1	27	3366721				3366721
13			1 Total		3366721				3366721
14		14 Total			3366721				3366721
15		18	59	15		5254981			5254981
16				16				5246789	5246789
17				17				5247117	5247117
18			59 Total			5254981		10493906	15748887
19		18 Total				5254981		10493906	15748887
20	24 Total				8621702	5254981		10493906	24370589

Figure 11, Shows the rollup operation in the data cube shown in figure 10.

Figure 7 shows part of the Web log cube. This figure shows part of the cube (excel pivot table) with 3 dimensions: DateTime, User and Protocol, where DateTime is at level Day. User and Protocol are in numeric codes (UserID and ProtocolID). For example ProtocolID 1 indicates protocol HTTP/1.1 method GET and server status 200. Server status 200 indicates a successful request, while 403 and 404 are code for request for forbidden URL and requested URL was not found, respectively.

Figure 8 shows the dicing operation on the cube shown in figure 7. This diced cube contains only two dimensions day and protocolID. Figure 9 is a slicing operation performed on the diced data cube. Figure 9 is a graphical representation of the diced cube shown in figure 8. As shown in the figure 9, we can see that no. of bytes transferred on day 23 was very few, which slightly increased on day 24 and 25 and on day 26 it was highest then decreased rapidly. Protocol 7 is used in big data transfers while in most of the data transfer user prefers protocol 1. Graphical representation make easy to perceive the mined data.



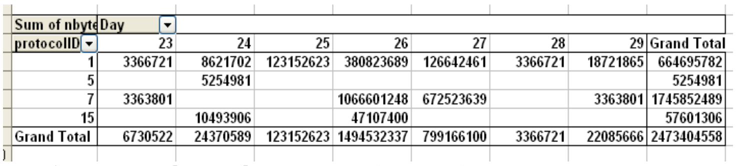
Sum of nbyte				protocolID			
Day	Hour	Minute	Seconds	1	5	15	Grand Total
24	13	24	42	5254981			5254981
		24 Total		5254981			5254981
	13 Total			5254981			5254981
	14	1	27	3366721			3366721
		1 Total		3366721			3366721
	14 Total			3366721			3366721
	18	59	15		5254981		5254981
			16			5246789	5246789
			17			5247117	5247117
		59 Total			5254981	10493906	15748887
	18 Total				5254981	10493906	15748887
24 Total				8621702	5254981	10493906	24370589
Grand Total				8621702	5254981	10493906	24370589

Figure 12, Slicing, Data cube on date-time dimension for day 24.

Figure 10, shows the Drill down operations. Here we Drill down the data cube shown in figure 8 into Hours, Minutes and seconds in the time dimension. Figure 11 illustrates the rollup operation over data cube shown in figure 10.

Using slicing operation (as shown in the figure 12) we can focus on the values of the specific cells. In the figure 12 we sliced the data cube in figure 10 for day 24. We can easily see the sum of the byte transferred of day 24 of each hour, minute and second.

Figure 13 shows the pivoting operation. In this example we rotate the two dimensions clock wise as a result the protocolID is now at the y axis and day are at the x axis as shown in the figure 13.



Sum of nbyte	Day							
protocolID	23	24	25	26	27	28	29	Grand Total
1	3366721	8621702	123152623	380823689	126642461	3366721	18721865	664695782
5		5254981						5254981
7	3363801			1066601248	672523639		3363801	1745852489
15		10493906		47107400				57601306
Grand Total	6730522	24370589	123152623	1494532337	799166100	3366721	22085666	2473404558

Figure 13, Shows the pivoting (rotation) operation where protocolID is now at the y-axis and days are at the x-axis.

Slicing, Dicing, Rollup, Drilldown and pivoting operations on a data cube provides great flexibility to an analyst to analyze the mined pattern in different perspective and at the different level of granularity.

We can automate these tasks by developing an interface which provides the possible options and display in the statistical terms or graphically. In excel this automation can be achieved using macro recording.

In the example discussed above we performed OLAM for Web Traffic Analysis. Similarly by changing the dimensions in the data cube we can also perform other analysis such as transition analysis, trend analysis, frequent access pattern etc.

5. Strength and Weakness of OLAM approach

The capability of OLAP to provide multiple and dynamic views of summarized data in a data warehouse sets a solid foundation for successful data mining. Moreover, data mining should be a human-centered process which does not generate patterns and knowledge automatically but allows user to interact with the system to perform exploratory data analysis. OLAM sets a good example for interactive data analysis and provides the necessary preparations for exploratory data mining.

The disadvantage of OLAP mining is that it still affected by the limitation of log files. In the current scenario these log files have very high data volumes which require implementation of such a Web access analysis engine that can support the high data volumes. Unfortunately, there is several performance and functionality problems that must be addressed before such a web access analysis engine can be implemented.

One such problem is how to handle the processing of very large, very sparse data cubes. Web access analysis introduces a number of fine-grained dimensions (such as seconds in the time dimension: day->hour->minute->second) that result in very large, very sparse data cubes. These very large, very sparse data cubes pose serious scalability and performance challenges to data aggregation and analysis, and more fundamentally, to the use of OLAP for such applications. While OLAP servers generally store sparse data cubes quite efficiently, OLAP servers generally do not roll-up these sparse data cubes very efficiently. For example, a newspaper Web site received 1.5 million hits a week against pages that contained articles on various subjects. The newspaper wanted to profile the behavior of visitors from each originating site at different times of the day, including their interest in particular subjects and which referring sites they were clicking through. The data is modeled by using four dimensions: ip address of the originating site (48,128 values), referring site (10,432 values), subject url (18,085 values), and hours of day (24 values). The resulting cube contains over 200 trillion cells, indicating clearly that the cube is extremely sparse. Each of the dimensions participates in a 2-level or 3-level hierarchy. To rollup such a cube along these dimension hierarchies by using the regular rollup operation supported by the OLAP server requires an estimated 10,000 hours (i.e. more than one year) on a single Unix server. As can be appreciated, the processing time required is unacceptable for the application. Accordingly, mechanisms are desired that can efficiently summarize data without having to roll-up sparse data cubes.

Other than this in OLAP a large mount of information such as how the user filled the tasks, the intension of the users and so on are missing. Some other limitations of this approach are:

- Non symmetric treatment of measure and dimension attributes.
- No navigation approaches for complex hierarchies and exploring multiple data cubes.
- Inability to provide a multiscale view.
- Standard Visualization tools scale poorly.

6. Conclusion

The huge development in the technology provides a great boost to the database and information industry, and makes a large number of databases and information repositories available for transaction management, information retrieval, and data analysis. “Repository of multiple heterogeneous data sources organized under a unified schema at a single site in order to facilitate management decision making“, is known as Data warehouse. Data warehouse is a wonderful resource for data mining. In descriptive level data mining data are presented in multidimensional data cubes and traditional analysis and reporting tools that are provided by OLAP techniques are used. It is known as OLAM (On Line Analytical Mining).

The users’ accesses to the web pages are recorded into a file called web logs. Web logs provide a huge repository of page access information which when mined properly can help improve system design, better marketing decisions etc. If data in web logs can be represented in the form of data warehouse (data cube) we can apply the OLAM techniques to analyze web usage patterns. In this chapter we tried to represent the analogy between the OLAM on normal database and OLAM on web logs. Using the example we try to prove that what actions/operations we can perform on simple data can also be performed on the web usage data. We first explained how is data preprocessing accomplished? Then how warehouse is implemented using multidimensional data cube? And different operations on data cube with the help of example. Then we presented the analogy for web usage data. We avoid the elimination because each data is important for us. We represented the concept hierarchy for web logs that help to perform various OLAP operations. Which type of patterns can be mined by using these operations are explained with the help of examples? Each technique has its own strength and weaknesses. The biggest weakness is huge web log results in sparse data cube which poses serious scalability and performance challenges.

7. References

1 U. Fayyad, G. Piatetsky-shapiro, and P. Smyth. From Data Mining to Knowledge Discovery: An Overview. In Advances in Knowledge Discovery and Data Mining. G. Piatetsky-Shapiro and J. Frawley, editors, AAAI Press, Menlo Park. C.A.. 1996.

2 S. Chaudhuri and U. Dayal, “An Overview of Data Warehousing and OLAP Technology”, ACM SIGMOD Record, Month? 1997. pp. 65-74.

3 O. R. Zaiane, M. Xin, J. Han. Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs. In Proceedings of Advances in Digital Libraries Conference (ADL 98), pages19-29, Santa Barbara, CA, USA, April 1998.

4 S. Sarawagi, R. Agrawal, and N. Megiddo. Discovery-driven Exploration of OLAP Data Cubes. In Proc. Of Extending Database Technology (EDBT98), pages 168-182, Valencta, Spain, March 1998.

5 J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web Usage Mining Discovery and application of Usage Patterns from Web Data. SIGKDD Exploration, 1[2]: 12-23, January 2000.

6 S. Sarawagi and G. Sathe. III: Intelligent, Interactive Investigation of OLAP Cubes. In Proc. Of the 2000 ACM SIGMOD International Conference on Management of Data, page 589, Dallas, Texas, USA, May 2000.

7 http://www.Data Warehouse and OLAP operations.htm.

8 Ralph Kimball “The Data Warehouse Toolkit”, John Willey and Sons, 1996.

9 Ralph Kimball, “Clicking with your Customer, Intelligence Enterprise”. Intelligent Enterprose, Jan05, 1999, Vol 2, No. 1.

10 The OLAP Report: http://www.olapreport.com/fasmi.htm release 2004.10.10.

11 Introduction to OLAP: http://www.dwreview.com/OLAP/Introduction_OLAP.html, release 2004.10.10.