High utility text mining using zealous algorithm

S.Thilagavathi
Department of Computer Science and Engineering
Muthayammal Engineering College
Namakkal, India

G.Sumathi
Department of Computer Science and Engineering
Muthayammal Engineering College
Namakkal, India

***Abstract*—Data mining is the process of analyzing data from different perspectives and summarizing it into useful information, Data mining uses the information from past data to analyze the outcome of a particular problem or situation that may arise. In this an important data mining task is to find the interesting patterns, and has a variety of applications, such as condition monitoring, cross marketing, and inventory prediction, where interestingness measures play an important role. With frequent pattern mining a pattern is regarded its occurrence frequency that exceeds a user specified threshold. The frequent pattern from a shopping transaction database refers to the discovery of sets of products that are frequently purchased together by customers. However, a user's interest may relate too many factors that are not necessarily expressed in terms of the occurrence frequency. To address the challenge, this paper proposes a new algorithm, data structure, for utility mining with the item set share framework. The proposed zealous algorithm contains some value for database query processing by keyword search. The related items retrieved from search engine to find high utility items.**

Keywords*— *Frequency, Utility, Pattern, Search, Request, Zealous algorithm, Query process

I. INTRODUCTION

A high utility pattern growth approach is the one to find high utility items without candidate generation because while the two-phase, candidate generation approach employed by prior algorithms first generates high TWU patterns with TWU being an interim, then anti-monotone measure and identifies high utility patterns from high TWU patterns, our approach directly discovers high utility patterns in a single phase without generating high TWU patterns.

Without recursive enumeration, a look ahead strategy is incorporated with our approach to identify the high utility patterns. Such a strategy is based on a closure property and a singleton property, and enhances the efficiency with dense data.

A linear data structure, CAUL, is proposed to represent original utility information or raw data in database, which targets the root cause with prior algorithms. That is, data structure to maintain the utility estimates instead of the original utility information, and thus can only determine the candidacy of a pattern but not the actual utility of the pattern in their first phase.

A. Data Mining Basic Analysis

In Data mining analysis, the extraction of hidden predictive information from large database, is a powerful new technology with great potential to help focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The data mining must be automated; prospective analyses offered the analyses of past events provided by retrospective tools typically decision support systems. Data mining tools can be used for too time consuming transaction applications.

To enhance the value of existing information resources, Data mining techniques can be implemented rapidly on existing software and hardware platforms and can be integrated with new systems in the platform of on-line, when implementation on high performance client/server or parallel processing computers. The profitable applications illustrate its relevance to business environment as well as a basic description of how data warehouse architectures can evolve to deliver the value of data mining to end users.

B. The Scope of Data Mining

In a large database, data mining derives its name from the similarities between searching for valuable business information— like finding linked products in gigabytes of store scanner data and information— and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities.

1) Automated prediction of trends and behaviors: Data mining is the process of finding predictive information in large databases. A predictive problem is targeted marketing to identify the targets. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting and other forms of default, and identifying segments.

2) Automated discovery of previously unknown patterns: Data mining tools search in a databases and identify previously hidden patterns in one step. The pattern discovery is the analysis of retail sales data to identify unrelated products that are often purchased together. Other pattern discovery problems include detecting frequent credit card transactions and identifying anomalous data that could represent data entry keying errors.

C. Architecture for Data Mining

The advanced techniques must be fully integrated with a data warehouse as well as flexible interactive business analysis tools. The data mining tools are currently operated outside of the warehouse, and it requires extra steps for extracting,

importing, and analyzing the data. Furthermore, The warehouse simplifies the application from data mining when new insights require operational implementation, integration. The data warehouse analysis process can be applied to improve business processes throughout the organization, in areas such as data management, fraud detection, new product rollout, and so on.

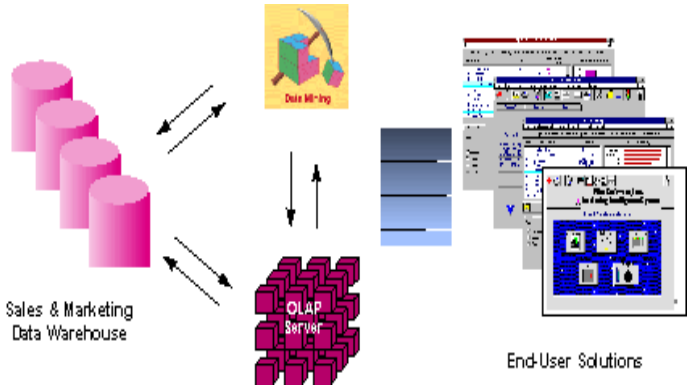


Fig. 1. Illustrates architecture for advanced analysis in a datawarehouse.

D. High-Utility Itemset

The high-utility item set mining is an extension of the problem of frequent pattern mining. A major task in data mining is frequent pattern mining, which obtains the problem is to find frequent patterns in transaction databases. Let me describe first the problem of frequent item set mining

Consider the following transaction database. A transaction database is a database which containing a set of transactions made by customers. A transaction is a set of process like items bought by the customers. For example, in the following database, the first customer bought items “a”, “b”, “c”, “d” and “e”, while the second one bought items “a”, “b” and “e”.

Transaction	items
T_1	{a, b, c, d, e}
T_2	{a, b, e}
T_3	{c, d, e}
T_4	{a, b, d, e}

The main objective of frequent item set mining is to find the frequent item sets. There are several algorithms has been proposed for solving this problem. These algorithms takes transaction database as input data and a parameter is called “minsup” means that minimum support threshold. All set of items that appears in minsup transactions will be returned by those algorithms.. For example, if we set minsup = 2, it will find several item sets such as the following.

Itemset	Support
{e}	4
{d, e}	3
{b, d, e}	2
{a}	3
...	...

Fig. 2. High Utility item set.

For example, consider the item set {b,d,e}. It have a support of 3 because it appears in three transactions, and it is said to be frequent because the support of {b,d,e} is no less than minsup.

E. Frequent Itemset Mining Limitations

The problem of frequent item set mining is major task in data mining process. When it comes to analyzing customer transactions, it has some important limitations. The first limitation is that purchase quantities are not taken into account. Thus, an item may only appear once or zero time in a transaction. If a customer has bought five breads, ten breads or twenty breads, it is viewed as the same. A second important limitation is that all items are viewed as having the same utility of weight.

The frequent pattern mining may also find many frequent patterns that are not interesting, so it is a problem of missing pattern. For example, one may find that {bread, milk} is a frequent pattern. However, from a business perspective, this pattern may be uninteresting because it does not generate much profit. Moreover, frequent pattern mining algorithms may miss the rare patterns that generate a high profit.

F. High-Utility Itemset Mining

To address these limitations, the problem of frequent item set mining has been redefined as the problem of high-utility item set mining.

transaction database with quantities

Trans.	items
T_0	a(1), b(5), c(1), d(3), (e,1)
T_1	b(4), c(3), d(3), e(1)
T_2	a(1), c(1), d(1)
T_3	a(2), c(6), e(2)
T_4	b(2), c(2), e(1)

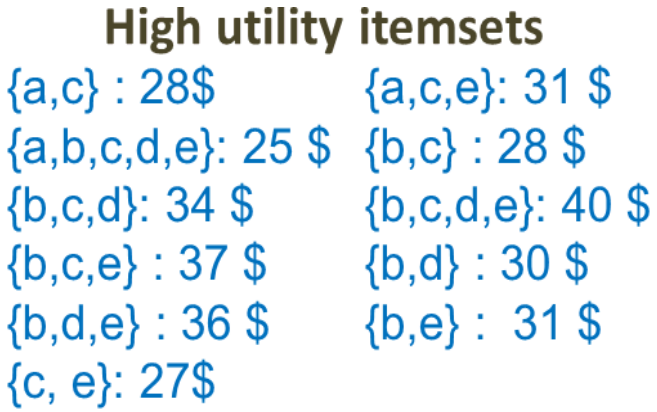
unit profit table

item	unit profit
a	5 $
b	2 $
c	1 $
d	2 $
e	3 $

Fig. 3. Transaction Database.

In high utility item set mining, a transaction database contains transactions where purchase quantities are taken into account as well as the unit profit of each item to purchase. For example, In the following transaction database. Consider the transaction T3. It indicates that the corresponding customer has bought two units of item “a”, six unit of item “c”, and two

units of item "e". Now the table on the right side indicates the unit profit of each item. For example, the unit profit of items "a", "b", "c", "d" and "e" are respectively 5$, 2$, 1$, 2$ and 3$. This means that each unit of "a" that is sold and generates a profit of 5$.



The item set {b,d} is considered to be a high-utility item set, because it has a utility of 40$. Which is no less than the minutil threshold that has been set to 25$ by the user. The high utility of an item set is calculated by using this method. In general the utility of an item set in a transaction is the quantity of each item from the item set multiplied by their unit profit.

G. *Zealous Algorithm*

The Zealous algorithm uses a two phase framework to find the frequent items in the log. The Zealous algorithm uses two threshold values to discover the frequent items. The log is act as a transaction database of search engine. The first threshold value is set based on the number of user contributions in the log. The main objective of Zealous algorithm is to find the frequent items in the log. This algorithm has applied to a sample search log collected from a local search engine to find the items in the log like keywords and URL values.

The log contained entities with its user. It is suitable for large database. The Zealous algorithm works with two threshold value on the log. In the two phase framework the keywords are passed to filtration process then these keywords are identified as frequent keywords. By using this two threshold values it identify the frequent URL clicks, otherwise the infrequent keywords to be leave in the log by zealous algorithm. The challenge task is to fix the threshold values, but in a search log, there will be several infrequent items. The infrequent item has no possibility of revealing a user's identity and it has to be published.

H. *Search Logs*

The Search engine such as Bing, Yahoo has log interactions with their users. A new entry is added to the search log, when a user submits a query and click on URL link. A search log has some schema as follows.

(USED-ID, QUERY, TIME, CLICKS)

An essential part of data mining research domain and frequent pattern mining is association rule mining. The mining process of finding the frequent patterns is based on the Apriori approach, which required more number of databases.

II. LITERATURE REVIEW

A. *Association Rules mining between Sets of Items in Large Databases: KaipingXue, Peilin Hong*

This proposed system has a large database of customer transactions. Each transaction consists of items purchased by a customer in a visit. An efficient algorithm that generates all significant association rules between items in the database. The algorithm incorporates buffer management and novel estimation and pruning techniques. To Applying this algorithm to present the result to sales data obtained from a large retailing company, which shows the effectiveness of the algorithm.

These algorithms to find all the rules relating items located on shelves A and B in the store. These rules may help shelf planning by determining if the sale of items on shelf A is related to the sale of items on shelf B. Discovering rules from data and its relation is based on different perspective. The purpose of quantitative rule is to automate the discovery of numeric laws of the scientific data. Too many formulas might be given for the data, because the algorithms and formulas should be based on the knowledge of the domain.

In different fields, each and every field the algorithm is run once to find rules. To finding all rules, it must make as many passes over the data as the number of combinations of attributes in the antecedent, which is exponentially large.

B. *Fast Algorithms for Mining Association Rules for finding frequent items: I.L'm, S. Szebeni, and L. Butty'n*

The problem of association rules is to find the rules between items in a large database of sales transactions. This system proposes two new algorithms for solving this problem that are fundamentally different from the known algorithms. Empirical evaluation shows that these algorithms to be different from the known algorithms by factors ranging from small problems to more than an order of magnitude for large problems. Besides the problem of discovering association rules, some other problems include the enhancement of the database capability with classification queries and similarity queries over time sequences.

C. *Efficient Algorithm for Finding High Utility Itemsets from Large Transactional Databases Using R-Hashing Technique: P. Tysowski and M. Hasan*

Generally the association rule mining is used to find the frequent item sets in large database. The Utility mining emerges that to mine the high utility item sets from databases which refers to finding the item sets with high profits in the data mining field. A challenging problem for the mining performance is it makes the huge number of high utility item sets, due to generating more potential high utility item sets. So it consumes higher process in large database and decreases the mining efficiency.

In existing system to lead high potential I/O operations the random memory allocation is used to store the candidate. Using this approach time consuming can be done and requires high memory space. In order to solve this problem in proposed system, it used sorting with R-hashing technique for the memory allocation. Then the candidate items are stored with their respective memory in UP tree.

A high-utility item set mining model was defined by Hamilton and it is a generalization of the share mining model. The main objective of high utility item set mining process is that to find all item sets that give utility greater or equal to the user specified threshold value. So setting the threshold value is more challenging task on user.

D. Efficient Algorithm for Mining Utility-Frequent Item sets: Y. Kim, A. Perrig, and G. Tsudik

The fast mining algorithms has, 2P-UF utility-frequent item set mining algorithm it is used to find all utility-frequent items. Due to the monotone property of the support measure it has a few disadvantages which are that unusable for mining of large datasets and less efficiency. The reversed way of candidate generation is the problem of algorithm that gives short item sets has large probability to be utility-frequent come at the very end. This candidate generation function is also slow and inefficient as it computes intersection of every pair of candidates in iteration.

Also the computation of support measure is inefficient because special data structures cannot be used for further process and it might be scan database once for every candidate. Finally, the two-phase algorithms is space consuming because all the utility-frequent candidates can be stored from the first phase, then in the second phase it is going to take for filter process. By merging both phases it is possible to avoid the wastage of space.

E. Constructing Up-Tree to Decrease Global Node Utilities: K.Kumar, Dr.V.Sumathy and J.Nafeesa Begum

In the mining processes of high utility item set mining, the tree-based framework applies the divide-and-conquer technique for frequent patterns. Thus, the search space can be divided into smaller subspaces for further process. The utilities of the nodes in the tree that are closer to the root of a global UP-Tree are reduced by applying strategy DGN. DGN is especially suitable for the databases containing lots of long transactions. But the problem is more items a transaction contains database, and its utilities can be discarded by DGN.

The database scanning process performed with the construction of a global UP-Tree structure. In the first scan process, for each transaction's TU is computed. The transactions are reorganized by pruning technique the unpromising items sorting by the remaining promising items in a fixed order. Any ordering can be used such as the lexicographic, or TWU order for sorting the remaining items. Each transaction after the process can be done by reorganization is called a reorganized transaction. However, the best ordering process of items is TWU descending order and its performance also good. During the construction of a global UP-Tree a function is going to process that Insert Reorganized Transaction. Memory management and number of searching level are high as compared with proposed system.

III. SYSTEM INFORMATION

A. Existing System

In data mining technology the important and challenging process is utility mining. It have some challenging problems, among utility mining problems utility mining with the item set share framework is a hard one as hold no interestingness measure. Interestingness measure is knowledge based one on shopping transaction. Prior problem on this is all the transactions incorporate with two-phase, and candidate generation approach with one exception that is however inefficient and not scalable with large databases.

When deals with large database, the two-phase approach suffers from scalability issue due to the huge number of candidates. Generally most of the prior utility mining algorithms with the item set share framework process with a two-phase, candidate generation approach. In the first phase find candidates of high utility patterns. And then scan the raw data one more time to identify high utility patterns from the candidates in the second phase.

When the process deals with huge number of candidates, it might be under the scalability and efficiency bottleneck. To reduce the number of candidates generated in the first phase has taken more effort, the challenge still persists when the raw data contains the minimum utility threshold is small or many long transaction. Such a huge number of candidates causes scalability issue not only in the first phase but also in the second phase, and consequently degrades the efficiency.

Drawbacks

- Inefficient and not scalable with large database
- Setting the framework is very hard in two face mining technologies
- Lack of strong pruning techniques and scalability

B. Proposed System

The proposed system has a new algorithm that find's the high utility pattern in a single phase without generating candidates. By a closure property and a singleton property a strategy is implemented to identify high utility patterns without enumeration. The linear data structure helps to compute a tight bound for powerful pruning and to directly identify high utility patterns in an efficient and scalable way, which focuses on the root cause problem with prior algorithms. The proposed system using zealous algorithm process with database query processing by some values which are posted on the formula.

A high utility pattern growth approach is proposed without candidate generation, because while in the two-phase, candidate generation approach employed by prior algorithms. First generates high TWU patterns with anti-monotone measure and then identifies high utility patterns from high TWU patterns, But the proposed approach directly find high utility patterns in a single phase without high TWU patterns.

The strength of this proposed approach comes from powerful pruning techniques based on tight upper bounds on utilities. This approach is incorporated with a look ahead strategy, which tries to identify high utility patterns earlier without recursive enumeration. Such a strategy is generally based on a closure property and a singleton property, and improves the efficiency with large data.

To represent the original utility information from raw data, a linear data structure CAUL is proposed, which targets the root cause with prior algorithms. This approach has a data structure to maintain the utility estimates instead of the original utility information.

The Search Log Algorithm called Zealous implemented for real time on-line problems, such as a server is allowed to decide whether to serve a request or not, and it can even wait idle. Basically an algorithm should decide to wait instead of serving ending requests, but consider the following case: the server is in the origin, and the only request released from its current position therefore it may have additional request then it might be wait. The real time implementation problem combines with moving a server could damage the quality of the overall service. The other real time problems like scheduling, it could be more dangerous to move the server for an isolated request, while completion time is objective. But the proposed zealous algorithm performance on basis on the keywords, queries, clicks and query pairs.

Understanding the information system design, interface development, and devising the information architecture for content collections or more important for information-searching process of online searcher. The web search transaction log analysis used for view the search.

A methodology is employed it consist of three stages, which are collection, preparation, and analysis. The strengths and limitations of transaction log analysis is trace out by represent this methodology.

Advantages

- Good and more efficient and scalable.
- Easy to mining the large database and get results in single phase.
- Save time

IV. SYSTEM DESIGN

Module Descriptions

A. User Registration

New Users have to register their details with the site at once. After completing the successful registration process, users are allowed to access the search details. That user details such as username, address, contact number, city, mail id, and password etc…Registration is one of the primary and necessary modules in any data management system. A registered user is an authorized user of a website, program, or other system which gives confidential. Registered users normally identify by previously entered details such as a username or e-mail address, and a password. System proves the identity by using logging process.

Registration module necessarily provides more personal information to a system from the customer side. The system can distinguish a logged-in user from other users by these information, and might use this property to store a history of users' actions or activity, possibly without their knowledge.

B. Categorical Word Search

The categorical word search module explains how the search terms are processed by the registered users. By using some keywords that user can enter, users will get the required information's from the search engines. These keywords may be anything and are frequently used to search, and the keywords related topic also retrieved from the search engine.

In the search module users can search for specific topic content on the particular site. The authorized user can search both for users details like location and for particular words. The content tab of Search, able to search for words appearing in the default rendering of node content on the site, which would include the default rendering of any CCK fields, Location fields, Taxonomy, etc.

C. Admin Login

The admin login module the admin login the site for maintain the product details, calculate the frequent and infrequent items list and analysis the graphs. An admin database has a set of procedures and tools to store and retrieve the information about product transaction. The database itself all the information stored about purchase.

The type of information stored in the database is defined by the corresponding data structures. The database structure consists of the tables, the relations between data, and domain. All the information's in the database to be stored in the form of tables. A database consists not only the data, also includes forms, queries and reports of transaction.

D. Add Product

In this module the process includes, the admin have to enter the product details on the database such as product -id, name, features, images and etc. And that information stored in database is maintained by admin. A DB consists of more than just the data, such as its related name, features, range. Database structure requires different views of the database from which to work. The representation of the database could be columns and rows, and then each and every product details can be placed on the top of the database.

E. View Frequent And Infrequent Item List

This is an algorithm module, this module includes the Zealous algorithm features, process explanation, and how the algorithm will get executed. This Zealous algorithm contains some values which are posted on the formula for database query processing. Zealous algorithm based on this the admin can view the frequent and infrequent products details.

Frequent Item sets are the recent techniques for characterizing the data. This problem is often viewed as association rules, although it is a more complex characterization of data, fundamentally on the discovery of frequent item sets. In this paper describe an approach which combines paramedical trees with association rule mining to discover infrequent patterns in data streams as well as any associations between infrequent patterns across multiple data streams. The two major issues in infrequent pattern mining is scalability in terms of memory requirements and pattern selection over time horizons that vary in span.

F. Graph

In this module the represented of the data and its relation by graph structure. Graphs will show the frequent and infrequent items to the user in a static format. With the use of graphs users will get the results in a clear view and easily can understand. Zealous graphs are defined to show the output on run time and to display the result values in a table view or a sequence of output.

The proposed graph structure captures only those item sets that need to define in a dataset. The dataset to be splinted into sub matrix, it perform based on the divide and conquer.

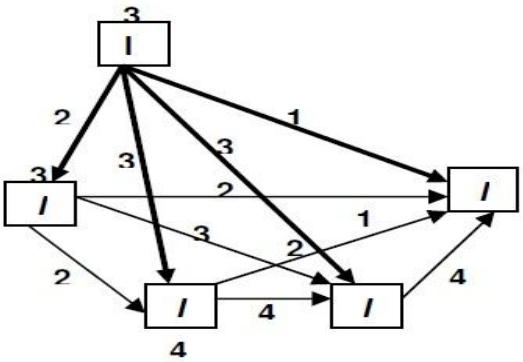


Fig. 4. Directed Graph of Database.

G. Queries

In this module the process might be done by user to the database admin. This process is based on query as request and response, the user used to send the queries to admin by customer. After visit this site if any doubt about product means send queries to admin and view reply.

A query is a request of customer for information from a database as reply. In order to retrieve information using the data objects and relationships in a data model, create a query and run it. The results returned by the database for a given query are known as a results set. The first step in creating a query is to select attributes from at least one data object, and then based on the relation it will process.

H. Queries Reply

This module is used to view the user queries and then sends the reply about the queries by admin. The query analyzer to analyze the user query request then based the optimization process the query response sends to customer.

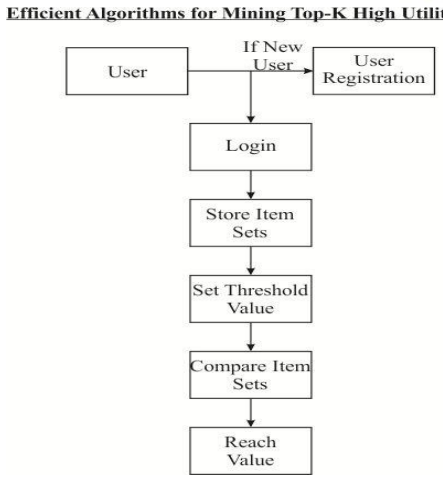


Fig. 5. Efficient Algorithm for Mining Top-K High Utility Itemsets.

I. Feedback

This module is used to view user feedback for analysis the product status and calculates the infrequent items list and then give the suggestion based on the user feed backs. By a graph representation, the infrequent items can be specified.

V. System Implementation

A. .NET

.NET is a "Software Platform". The components that make the process of .NET platform are collectively called the .NET Framework. The .NET Framework is a managed, type-safe environment for developing and executing applications. The .NET Framework manages all aspects of program execution, like, allocation of memory for the storage of data and instructions, granting and denying permissions to the application, managing execution of the application and reallocation of memory for resources that are not needed.

Components of .NET Framework:

- Common Language Runtime (CLR)
- Class Libraries.

B. Common Language Runtime (CLR)

The CLR is described as the "execution engine" of .NET. It provides the environment which the programs run and it includes program coding, compilation. CLR manages the execution of programs and provides core services, such as code compilation, memory allocation, thread management, and garbage collection.

C. Common Language Specification (CLS)

Class library is the second major entity of the .NET Framework which is designed to integrate with the common language runtime. CLS specify the common feature of different languages.

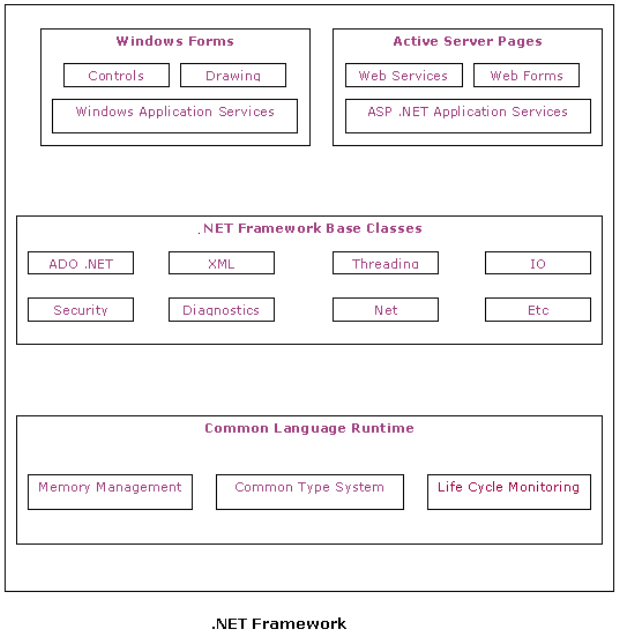


Fig. 6. DOT NET Framework.

D. ASP.NET

ASP.NET is a programming framework used to create enterprise-class Web Applications. These applications are accessible on a global database system to efficient information management. The DOTNET Framework used for distributed software with Internet functionality and interoperability. It includes class libraries, much language support.

Advantages of ASP.NET

- Reduces the amount of code
- Simpler and easier to maintain
- Easy to write and maintain because the source code and HTML are together.
- Easy deployment

E. Client And Server Scripting

JavaScript and VBScript generally used for Client-side scripting. A client-side script runs only on a browser that supports scripting and specifically the scripting language. It uses the simplified language HTML, on the other side ASP.NET is executed on server side. Through the browser client and server can communicate.

F. MS-SQL Server 2005

MS SQL Server is an efficient database management system and it supports GUI features and all programming language support, Visual Studio Application which can be used to develop effective and rich applications.

- SQL Profiler: Optimizing the database with performance issues.
- Service Manager: It is used to control the MS-SQL Server
- Data Transformation Services (DTS): Importing exporting data

G. SQL Server Architecture

The data in a database is organized into the logical component that is visible to users. A database has the logical components such as tables, views, procedures, and users. The physical implementation of files is largely transparent. The database administrator only needs to work with the physical implementation.

The SQL Server database engine allows multiple users to access the databases on a server. Each instance of SQL Server makes all databases in the instance available to all users that connect to the instance, with the defined security permissions. The connection is associated with a particular database on the server. This database is called the current database.

CONCLUSION AND FUTURE WORK

The proposed system has a new zealous algorithm for utility mining, with the item set share framework on the transaction database. In this system using the algorithms and frequent mining find high utility patterns without candidate generation.

Future works includes:

- A linear data structure, CAUL is proposed, which targets the root cause of the two phases, candidate generation approach adopted by prior algorithms, and data structures cannot keep the original utility information is an root cause.
- A high utility pattern growth approach is implemented, which integrates a pattern enumeration strategy, pruning by utility upper bounding, and CAUL.
- This approach is enhanced significantly by the look ahead strategy that identifies high utility patterns without enumeration. In the future, the process is going to deals with high utility sequential pattern mining, parallel and distributed algorithms, and its application in big data analytics.

REFERENCES

[1] R. Agarwal, C. Aggarwal, and V. Prasad, "Depth first generation of long patterns," in Proc. ACM SIGKDD Int. Conf. Knowledge. Discovery Data Mining, 2000, pp. 108–118.

[2] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 1993, pp. 207–216.

[3] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. 20th Int. Conf. Very Large Databases, 1994, pp. 487–499.

[4] C. F. Ahmed, S. K. Tanbeer, B.-S. Jeong, and Y.-K. Lee, "Efficient tree structures for high utility pattern mining in incremental databases," IEEE Trans. Knowl. Data Eng., vol. 21, no. 12, pp. 1708–1721, Dec. 2009.

[5] R. Bayardo and R. Agrawal , "Mining the most interesting rules," in Proc. 5th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 1999, pp. 145–154.

[6] F. Bonchi, F. Giannotti, A. Mazzanti, and D. Pedreschi, "ExAnte: A preprocessing method for frequent-pattern mining," IEEE Intell. Syst., vol. 20, no. 3, pp. 25–31, May/Jun. 2005.

[7] F. Bonchi and B. Goethals, "FP-Bonsai: The art of growing and pruning small FP-trees," in Proc. 8th Pacific-Asia Conf. Adv. Knowledge. Discovery Data Mining, 2004, pp. 155–160.

[8] F. Bonchi and C. Lucchese, "Extending the state-of-the-art of constraint-based pattern discovery," Data Knowledge .Eng., vol. 60, no. 2, pp. 377–399, 2007.

[9] C. Bucila, J. Gehrke, D. Kifer, and W. M. White, "Dualminer: A dual-pruning algorithm for itemsets with constraints," Data Mining Knowledge. Discovery, vol. 7, no. 3, pp. 241–272, 2003.

[10] C. H. Cai, A. W. C. Fu, C. H. Cheng, and W. W. Kwong, "Mining association rules with weighted items," in Proc. Int. Database Eng. Appl. Symp., 1998, pp. 68–77.

[11] R. Chan, Q. Yang, and Y. Shen, "Mining high utility item sets," in Proc. Int. Conf. Data Mining, 2003, pp. 19–26.

[12] J. R. Ullmann, (1976) An algorithm for sub graph isomorphism. J. ACM, 23, pp. 31-42.

[13] D. J. Cook and L. B. Holder, (1994) Substructure discovery using minimum description lengthand background knowledge, Journal of Artificial intelligence Research, 1, 231-255.

[14] Holder, L. B. Holder, Cook, D. J. Cook, Djoko, S. Djoko, (1994) Substructure Discovery in the SUBDUE system, In Proc. AAAI'94 Workshop knowledge Discovery in Databases (KDD'94), pp 169-180.

[15] S.J. Yen and A.L.P. Chen. (1996) An Efficient Approach to Discovering Knowledge. In Proc. Of the IEEE/ACM International Conference on Parallel and Distributed Information Systems, Pages 8-18.

[16] A. Inokuchi, T. Washio, H. Motoda, (1998) An Apriori-based Algorithm for Mining Frequents substructures from Graph Data. In proc. 2000 European Symp. Principle of Data mining and knowledge Discovery (PKDD'00), pp. 13-23.

[17] Y.J. Li, C. Luo, and S.M. Chung, "Text Clustering with Feature Selection by Using Statistical Data," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 5, pp. 641-652, May 2008.

[18] W. Lian, D.W.-L. Cheung, N. Mamoulis, and S.-M. Yiu, "An Efficient and Scalable Algorithm for Clustering XML Documents by Structure," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 1,pp. 82-96, Jan. 2004.

[19] Z. Liu and Y. Chen, "Identifying Meaningful Return Information for XML Keyword Search," Proc. ACM SIGMOD Int'l Conf.Management of Data, 2007.

[20] Holder, L. B. Holder, Cook, D. J. Cook, Djoko, S. Djoko, (1994) Substructure Discovery in the SUBDUE system, In Proc. AAAI'94 Workshop knowledge Discovery in Databases (KDD'94), pp 169-180.

[21] D. J. Cook and L. B. Holder, (1994) Substructure discovery using minimum description lengthand background knowledge, Journal of Artificial intelligence Research, 1, 231-255.

[22] Holder, L. B. Holder, Cook, D. J. Cook, Djoko, S. Djoko, (1994) Substructure Discovery in the SUBDUE system, In Proc. AAAI'94 Workshop knowledge Discovery in Databases (KDD'94), pp 169-180.

[23] S.J. Yen and A.L.P. Chen. (1996) An approach to Discovering Knowledge from Large Databases. In Proc. Of the IEEE/ACM International Conference on Parallel and Distributed Information Systems, Pages 8-18.

[24] P. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, 1st ed. Addison Wesley, May 2005.

[25] A. Brandi and P. Blonde, "A survey of fuzzy clustering algorithms for patternt recognition." IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 29, no. 6, pp. 778–785, Dec. 1999.

[26] A. Silic, M.-F. Moans, L. Zmak, and B. Basic, "Comparing document classification schemes using k-means clustering," vol. 5177, pp. 615–624, 2008.