*ISSN: 2278 – 1323*
*International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*
*Volume 1, Issue 7, September 2012*

# Mining on Medicine Data

*AJAY KUSHWAHA*

*MANOJEET ROY*

***ABSTRACT: In this paper, we propose an approach for Mining as well as data clustering, we have taken some medical data set we not only able to find out the disease detail as well as clustering. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with low distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Most data-mining methods assume data is in the form of a feature-vector (a single relational table) and cannot handle multi-relational data. Two fundamental issues regarding the effectiveness of information gathering from the Web: mismatch and overload. Mismatch means some useful and interesting data has been overlooked, whereas overload means some gathered data is not what users want. Classification and clustering has become an increasingly popular method of multivariate analysis over the past two decades, and with it has come a vast amount of published material. Since there is no journal devoted exclusively to cluster analysis as a general topic and since it has been used in many fields of study. Traditional techniques related to information retrieval (IR) have touched upon the fundamental issues***

***KEYWORDS:- clustering, nearest neighbor, reciprocal nearest neighbor, complete link, probabilistic analysis.***

## 1.INTRODUCTION

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies 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 automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations[1].

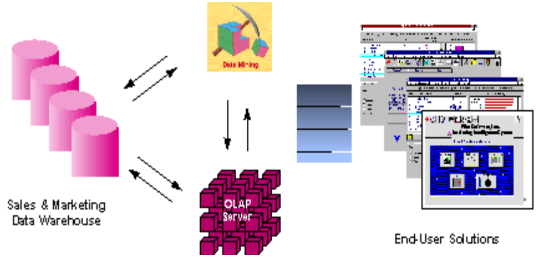


**Figure 1.1 - Integrated Data Mining Architecture**

Traditional techniques related to information retrieval (IR) have touched upon the fundamental issues. However, IR-based systems neither explicitly describe how the systems can act like users nor discover exotic knowledge from very large data sets to answer what users really want. This issue has challenged the artificial intelligence (AI) community to address "what has information gathering to do with AI" . For a short while, many intelligent agent based approaches have been grappling with this challenge. Unfortunately, agent-based approaches can only show us the architectures of information gathering systems. They cannot provide strategies for finding interesting and useful knowledge from data to overcome the fundamental issues. Web intelligence (WI) is a new direction which can provide a new approach to solve this problem. Currently, the application of data mining techniques to Web data, called Web mining, is used to discover patterns from data (e.g., user feedback or user log data).

Web information Web mining can be classified into four categories: Web usage, Web structure, Web content, and Web user profiles In this paper, we develop an mining technique to overcome the above drawbacks.

In the beginning, we assume that the training set only includes positive documents and that the system can discover some patterns from the training set. During the execution, the system might select a small amount of documents and require users to label them as either positive or negative (user feedback). We also assume that user interests learning and pattern recognition.

As the marketing director you have access to a lot of information about all of your customers: their age, sex, credit history and long distance calling usage. The good news is that you also have a lot of information about your prospective customers: their age, sex, credit history etc. Your problem is that you don't know the long distance calling usage of these prospects (since they are most likely now customers of your competition). You'd like to concentrate on those prospects who have large amounts of long distance usage. You can accomplish this by building a model. Table 2 illustrates the data used for building a model for new customer prospecting in a data warehouse.

**Table 1.1** - Data Mining for Prospecting

| information | Customers | Prospects |
|---|---|---|
| General information (e.g.demographic data) | Known | Known |
| Proprietary information | Known | Target |

The goal in prospecting is to make some calculated guesses about the information in the lower right hand quadrant based on the model that we build going from Customer General Information to Customer Proprietary Information. For instance, a simple model for a telecommunications company might be:

98% of my customers who make more than $60,000/year spend more than $80/month on long distance

This model could then be applied to the prospect data to try to tell something about the proprietary information that this telecommunications company does not currently have access to. With this model in hand new customers can be selectively targeted.

Test marketing is an excellent source of data for this kind of modeling. Mining the results of a test market representing a broad but relatively small sample of prospects can provide a foundation for identifying good prospects in the overall market. Table 3 shows another common scenario for building models: predict what is going to happen in the future.

**Table 1.2** - Data Mining for Predictions

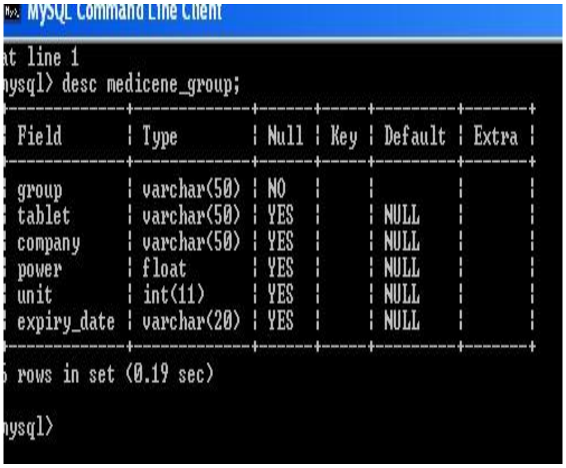
| Information | Yesterday | Today | Tomorrow |
|---|---|---|---|
| Static information and current plans (e.g. demographic data, marketing plans) | Known | Known | Known |
| Dynamic information (e.g. customer transactions) | Known | Known | Target |

If someone told you that he had a model that could predict customer usage how would you know if he really had a good model? The first thing you might try would be to ask him to apply his model to your customer base - where you already knew the answer. With data mining, the best way to accomplish this is by setting aside some of your data in a vault to isolate it from the mining process. Once the mining is complete, the results can be tested against the data held in the vault to confirm the model's validity. If the model works, its observations should hold for the vaulted data.

## 2.METHODOLGY

### 2.1 Medicine Group

Based on the general requirement we create the number of required data field for the different medicine specification ,some tool as been referred to create kind of database description for idea to create our own data set.



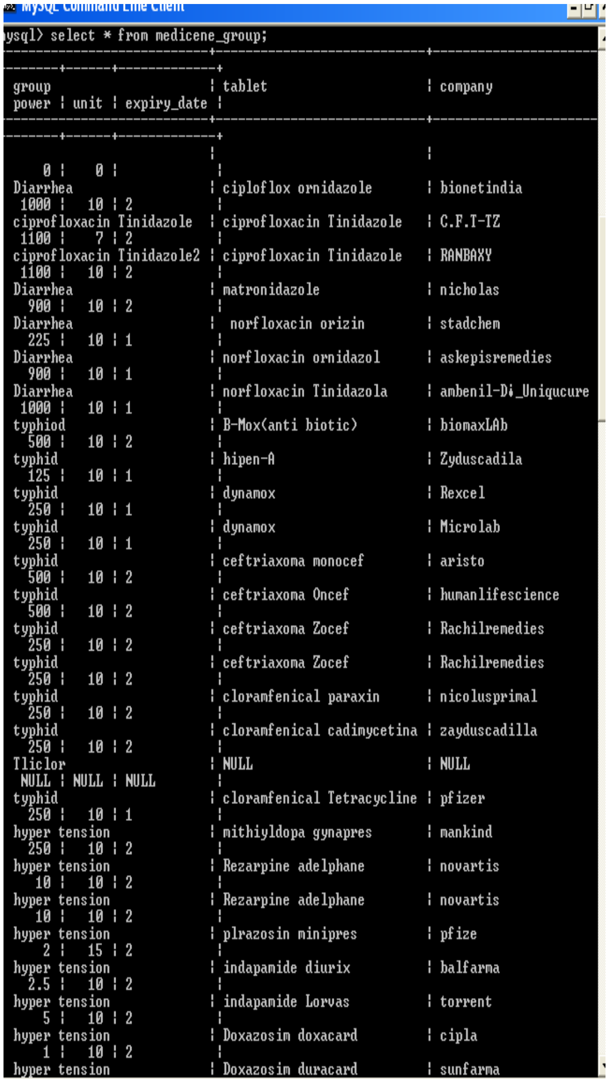
```
t line 1
ysql> desc medicene_group;
+-------------+-------------+------+-----+---------+-------+
| Field       | Type        | Null | Key | Default | Extra |
+-------------+-------------+------+-----+---------+-------+
| group       | varchar(50) | NO   |     |         |       |
| tablet      | varchar(50) | YES  |     | NULL    |       |
| company     | varchar(50) | YES  |     | NULL    |       |
| power       | float       | YES  |     | NULL    |       |
| unit        | int(11)     | YES  |     | NULL    |       |
| expiry_date | varchar(20) | YES  |     | NULL    |       |
+-------------+-------------+------+-----+---------+-------+
 rows in set (0.19 sec)

ysql>
```

**Figure 2.1:- Snap shot of structure of medicine group which is created**

- In addition to the classic "Web of documents" W3C is helping to build a technology stack to support a "Web of data," the sort of data you find in databases. The ultimate goal of the Web of data is to enable computers to do more useful work and to develop systems that can support trusted interactions over the network. The term "Semantic Web" refers to W3C's vision of the Web of linked data. Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data. Linked data are empowered by technologies RDF The Resource Description Framework (*RDF*) is a family of World Wide Web.

- Query languages go hand-in-hand with databases. If the Semantic Web is viewed as a global database, then it is easy to understand why one would need a query language for that data.



```
ysql> select * from medicene_group;
----------------------------+-----------------------------+--------------------
--------+------+-------------+
 group                      | tablet                      | company
 power | unit | expiry_date |
----------------------------+-----------------------------+--------------------
--------+------+-------------+
                            |                             |
     0 |    0 |             |
 Diarrhea                   | ciploflox ornidazole        | bionetindia
  1000 |   10 | 2           |
 ciprofloxacin Tinidazole   | ciprofloxacin Tinidazole    | C.F.T-TZ
  1100 |    7 | 2           |
 ciprofloxacin Tinidazole2  | ciprofloxacin Tinidazole    | RANBAXY
  1100 |   10 | 2           |
 Diarrhea                   | matronidazole               | nicholas
   900 |   10 | 2           |
 Diarrhea                   |  norfloxacin orizin         | stadchem
   225 |   10 | 1           |
 Diarrhea                   | norfloxacin ornidazol       | askepisremedies
   900 |   10 | 1           |
 Diarrhea                   | norfloxacin Tinidazola      | ambenil-D♦_Uniqucure
  1000 |   10 | 1           |
 typhiod                    | B-Mox(anti biotic)          | biomaxLAb
   500 |   10 | 2           |
 typhid                     | hipen-A                     | Zyduscadila
   125 |   10 | 1           |
 typhid                     | dynamox                     | Rexcel
   250 |   10 | 1           |
 typhid                     | dynamox                     | Microlab
   250 |   10 | 1           |
 typhid                     | ceftriaxoma monocef         | aristo
   500 |   10 | 2           |
 typhid                     | ceftriaxoma Oncef           | humanlifescience
   500 |   10 | 2           |
 typhid                     | ceftriaxoma Zocef           | Rachilremedies
   250 |   10 | 2           |
 typhid                     | ceftriaxoma Zocef           | Rachilremedies
   250 |   10 | 2           |
 typhid                     | cloramfenical paraxin       | nicolusprimal
   250 |   10 | 2           |
 typhid                     | cloramfenical cadimycetina  | zayduscadilla
   250 |   10 | 2           |
 Tliclor                    | NULL                        | NULL
  NULL | NULL | NULL        |
 typhid                     | cloramfenical Tetracycline  | pfizer
   250 |   10 | 1           |
 hyper tension              | mithiyldopa gynapres        | mankind
   250 |   10 | 2           |
 hyper tension              | Rezarpine adelphane         | novartis
    10 |   10 | 2           |
 hyper tension              | Rezarpine adelphane         | novartis
    10 |   10 | 2           |
 hyper tension              | plrazosin minipres          | pfize
     2 |   15 | 2           |
 hyper tension              | indapamide diurix           | balfarma
   2.5 |   10 | 2           |
 hyper tension              | indapamide Lorvas           | torrent
     5 |   10 | 2           |
 hyper tension              | Doxazosim doxacard          | cipla
     1 |   10 | 2           |
 hyper tension              | Doxazosim duracard          | sunfarma
```

**Fig 2.2:-Database creation**

Other data base required to create the basic symptoms of disease

Report can be generate by identifying all the requirement specification of the diseases for example in the third week of typhoid fever, a number of complications can occur:

- Intestinal hemorrhage due to bleeding in congested Payer's patches; this can be very serious but is usually not fatal.

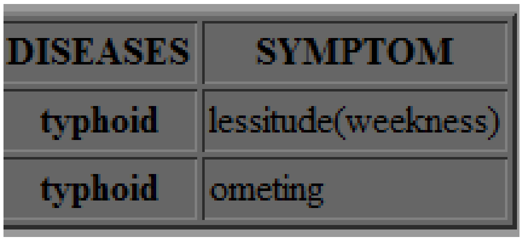
- Intestinal perforation in the distal ileum: this is a very serious complication and is frequently fatal. It may occur without alarming symptoms until septicemia or diffuse peritonitis sets in.

- Encephalitis

- Neuropsychiatric symptoms (described as "muttering delirium" or "coma vigil"), with picking at bedclothes or imaginary objects.

- Metastatic abscesses, cholecystitis, endocarditis and osteitis

The fever is still very high and oscillates very little over 24 hours. Dehydration ensues and the patient is delirious (typhoid state). By the end of third week the fever has started reducing this (defervescence). This carries on into the fourth and final week.



| DISEASES | SYMPTOM |
|---|---|
| typhoid | lessitude(weekness) |
| typhoid | ometing |

**Fig: 2.3:-Cluster and its formation**

Data clustering is a method in which we make cluster of objects that are somehow similar in characteristics. The criterion for checking the similarity is implementation dependent.

Clustering is often confused with classification, but there is some difference between the two. In classification the objects are assigned to pre defined classes, whereas in clustering the classes are also to be defined. Precisely, Data Clustering is a technique in which, the information that is logically similar is physically stored together. In order to increase the efficiency in the database systems the number of disk accesses are to be minimized. In clustering the objects of similar properties are placed in one class of objects and a single access to the disk makes the entire class available.

In this section some frequently used terms are defined.

### 2.1**Cluster**

A cluster is an ordered list of objects, which have some common characteristics. The objects belong to an interval [a , b], in our case [0 , 1] [1]

### 2.2 **Distance Between Two Clusters**

The distance between two clusters involves some or all elements of the two clusters. The clustering method determines how the distance should be computed. [1]

### 2.3 **Similarity**

A similarity measure SIMILAR ( $D_i$, $D_j$ ) can be used to represent the similarity between the documents. Typical similarity generates values of 0 for documents exhibiting no agreement among the assigned indexed terms, and 1 when perfect agreement is detected. Intermediate values are obtained for cases of partial agreement. [1]

### 2.4 **Average Similarity**

If the similarity measure is computed for all pairs of documents ( $D_i$, $D_j$ ) except when i=j, an average value AVERAGE SIMILARITY is obtainable. Specifically, AVERAGE SIMILARITY = CONSTANT SIMILAR ( $D_i$, $D_j$ ), where i=1,2,….n and

j=1,2,….n and i < > j

### 2.5 **Threshold**

The lowest possible input value of similarity required to join two objects in one cluster.

### 2.6 **Similarity Matrix**

Similarity between objects calculated by the function SIMILAR ($D_{i,}$,$D_j$), represented in the form of a matrix is called a similarity matrix.

The basic algorithm is very simple:

1. Start with each point in a cluster of its own

2. Until there is only one cluster

   (a) Find the closest pair of clusters

   (b) Merge them

This is only achieved by the algorithm which is design that is Begin with the disjoint clustering having level L(0) = 0 and sequence number

m = 0.

2. Find the least dissimilar pair of clusters in the current clustering, say pair

(r), (s), according to

d[(r),(s)] = min d[(i),(j)]

where the minimum is over all pairs of clusters in the current clustering. Increment the sequence number : m = m +1. Merge clusters (r) and (s) into a

ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 1, Issue 7, September 2012

single cluster to form the next clustering m. Set the level of this clustering to

L(m) = d[(r),(s)]

2. Update the proximity matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The proximity between the new cluster, denoted (r,s) and old cluster (k) is defined in this way:

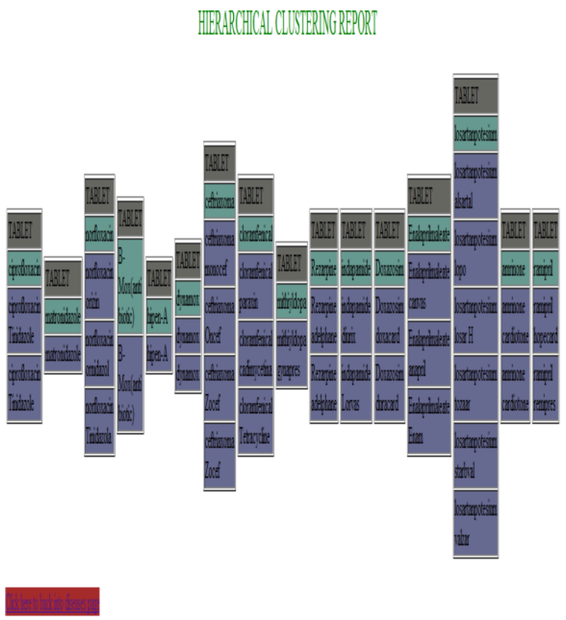
d[(k), (r,s)] = min d[(k),(r)], d[(k),(s)]

If all objects are in one cluster, stop. Else, go to step 2.

The number of match is only found by the minimum number of distance in between d[(r),(s)] = min d[(i),(j)],where clusters in the current clustering, say pair
(r), (s).

To obtain the goal we create cluster of various medicine based on needed information for the disease as well as required medicine group where various combination medicine genrally doctor refer for example ciplofloxin medicine come with its combinational medicine group that is tinidazola we form a cluster of this kind of group medicine

Figure of obtain cluster from the various combinational medicine is shown next similarly we get the report from the all information best on the mostly suggestion by the doctor .



HIERARCHICAL CLUSTERING REPORT

TABLET
ciprofloxacin
ciprofloxacin Tinidazole
ciprofloxacin Tinidazole

TABLET
matronidazole
matronidazole

TABLET
norfloxacin
norfloxacin orizin
norfloxacin ornidazol
norfloxacin Tinidazola

TABLET
B-Mox(anti biotic)
B-Mox(anti biotic)

TABLET
hipen-A
hipen-A

TABLET
dynamox
dynamox
dynamox

TABLET
ceftriaxoma
ceftriaxoma monocef
ceftriaxoma Oncef
ceftriaxoma Zocef
ceftriaxoma Zocef

TABLET
cloranfenical
cloranfenical paraxin
cloranfenical cadmycetna
cloranfenical Tetracycline

TABLET
mithyldopa
mithyldopa gynapres

TABLET
Rezarpine
Rezarpine adelphane
Rezarpine adelphane

TABLET
indapamide
indapamide dirix
indapamide Lorvas

TABLET
Doxazosin
Doxazosin doxacard
Doxazosin duracard

TABLET
Enalaprilmaleate
Enalaprilmaleate canvas
Enalaprilmaleate anapril
Enalaprilmaleate Enam

TABLET
losartanpotesium
losartanpotesium alsartal
losartanpotesium lopo
losartanpotesium losar H
losartanpotesium tozaar
losartanpotesium starboil
losartanpotesium valzar

TABLET
amrinone
amrinone cardiotone
amrinone cardiotone

TABLET
ramipril
ramipril hopecard
ramipril remipres

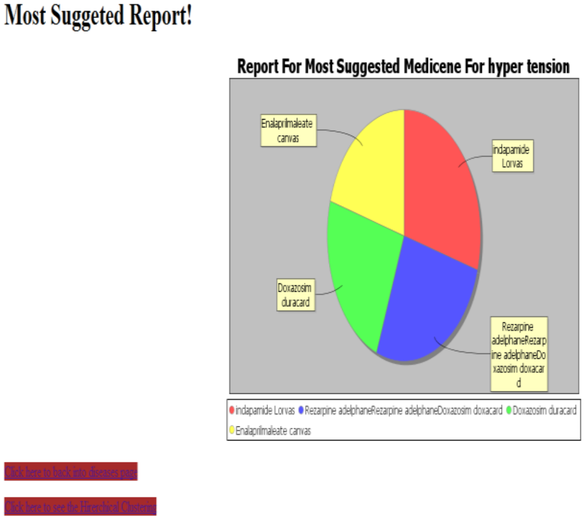
**Fig: 2.4:-Various combination of medicine cluster**

The cluster formation is obtain from the various medicine group ,in our this project we can create a report on the basis of the following points:-

- Based on the symptoms
- Based on the mostly suggested medicine by doctor .
- Based on the availability of the medicine in the market.

### 2.2 Report Generation

Over the years, we have often likened non-science-based medical belief systems to religion. It’s not a hard argument to make. Religion involves believing in things that can’t be proven scientifically; indeed, religion makes a virtue out of ignoring the evidence and accepting various beliefs on faith alone. Similarly, alternative medicine frequently tells you that you have to believe in the therapy, dedicate yourself completely to it, in order for it to work. Of course, as I've also mentioned before, it is that insistence on belief and total commitment shared by religion and alternative medicine that provides quacks with an "out" when their treatments don't yield the promised results, their frequent excuse being to blame the patient. He didn't believe hard enough. In a reverse of The Secret, which states that you can bring good things to yourself by simply wanting it, in alt-med world, it's all too often implied (or even more than implied).


Most Suggeted Report!

Report For Most Suggested Medicene For hyper tension
Enalaprilmaleate canvas
indapamide Lorvas
Doxazosim duracard
Rezarpine adelphaneRezarpine adelphaneDoxazosim doxacard
● indapamide Lorvas ● Rezarpine adelphaneRezarpine adelphaneDoxazosim doxacard ● Doxazosim duracard ● Enalaprilmaleate canvas

Click here to back into diseases page

Click here to see the Hirerchical Clustering

**Fig: 2.5:-most suggested medicine for hyper tension**

## CONCLUSION

The investigation carried out in this project leads to a very important conclusion. There is no doubt that numerous discovered patterns can be found from the Web data using data mining techniques. However, it is ineffective to use the discovered patterns in Web user profile mining due to the ambiguities in the data values (terms). The consequent result is that we obtain some inappropriate discovered patterns and many discovered patterns include uncertainties. In this paper, we develop an ontology mining technique to provide a solution for this challenge. A discovered ontology in this research consists of two parts: the top backbone and the base backbone. The former illustrates the linkage between compound classes of the mining. The latter illustrates the linkage between primitive classes and compound classes.

The project has referenced and discussed the issues on the specified algorithms for the data analysis. The analysis does not include missing records. The application can be used to demonstrate how data mining technique can be combined with medical data sets and can be effectively demonstrated in modifying the clinical research. This study clearly shows that data mining techniques are promising for clinical datasets. Our future work will be related to missing values and applying various algorithms for the fast implementation of records. In addition, the research would be focusing on spatial data clustering to develop a new spatial data mining technique. This can only be demonstrated by using medical data set where number of medicine and the disease symptoms are mapped as well as cluster of various combination medicines are formed ,this whole system we cannot consider it as artificial intelligent system. We get some certain outcome from presumptions.

This thesis study can be extended by the following subjects

- In this thesis study, various clustering method and mining concept
- Best way to find match in between object distance.
- This work can be extended with different algorithm which is work similar with other press describe algorithm with the less complicated successive steps
- Various other area like share market whether forecasting type of problem can also implement .

## REFERENCES

R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval.Addison Wesley, 1999.

[2] D.A. Grossman and O. Frieder Information Retrieval Algorithms and Heuristics. Kluwer Academic, 1998.

[3] M.N. Garofalakis, R. Rastogi, S. Seshadri, and K. Shim, ―Data Mining and the Web: Past, Present and Future,‖ Proc. ACM CIKM

[4] Int'l Workshop Web Information and Data Management, pp. 43-47,1999.

[5] K.S. Jones, ―Information Retrieval and Artificial Intelligence,‖ Artificial Intelligence, vol. 114, nos. 1-2, pp. 257-281, 1999.

[6] S.M. Madria, S.S. Bhowmick, W.K. Ng, and E.-P. Lim, ―ResearchIssues in Web Data Mining,‖ Proc. First Int'l Conf. Data Warehousingand Knowledge Discovery, pp. 303-312, 1999.

[7] S.K. Pal and V. Talwar, ―Web Mining in Soft ComputingFramework: Relevance, State of the Art and Future Directions,‖ IEEE Trans. Neural Networks, vol. 13, no. 5, pp. 1163-1177, 2002.

[8] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, ―Web Usage Mining: Discovery and Applications of Usage Pattern from Web Data,‖ SIGKDD Explorations, vol. 1, no. 2, pp. 12-23, 2002.

[9] N. Zhong, J. Liu, and Y.Y. Yao, ―In Search of the Wisdom Web,‖ Computer, vol. 35, no. 11, pp. 27-31, Nov. 2002.

[10] S. Tsumoto and S. Hirano, ―Visualization of Rule's Similarity Using Multidimensional Scaling,‖ Proc. Third IEEE Int'l Conf. Data Mining, pp. 339-346, 2003.

[11] T.Y. Yan, M. Jacobsen, H. Garcia-Molina, and U. Dayal, ―From User Access Patterns to Dynamic Hypertext Linking,‖ Proc. Fifth Int'l World Wide Web Conf., 1996.

**Prof. Ajay Kushwaha**, Reader c.s.e Deptt.RCET ,bhilai M.C.A , Mtech(CS),PhD (CSE) pursuing from CSVTU ,Chhattisgarh Research area – MANET, Address : RCET ,KOHKA - KURUD ROAD ,KOHKA , BHILAI -490023

**Manojeet Roy**, Mtech (Scholar) in Computer Technology in C.S.E. department from RCET Bhilai Chhattisgarh Under the CSVTU Universityroy. Address : RCET ,KOHKA - KURUD ROAD ,KOHKA , BHILAI -490023