Extraction of Interesting Rules from Internet Search Histories

Md. Asaduzzaman, Md. Shahjahan
Department of Electrical & Electronic Engineering,
Khulna University of Engineering & Technology (KUET), Khulna-9203
Email: mdjahan8@yahoo.com

Kazuyuki Murase
Department of Human & Artificial Intelligence Systems,
Graduate School of Engineering, University of Fukui,
3-9-1 Bunkyo, Fukui 910-8507, Japan. Email: murase@synapse.his.u-fukui.ac.jp

Abstract—This paper reports a method that finds out interesting rules from the heterogeneous Internet search histories. Rule extraction aims to improve business performance through an understanding of past and present search histories of customers. A challenging task is to determine interesting rules from their heterogeneous search histories of shopping in the Internet. Customers visit web pages one after another and leave their valuable search information behind. Firstly we produce a homogeneous data set from their heterogeneous search histories. It is difficult task to produce a homogeneous data from heterogeneous data without changing their characteristics of data. Secondly these data are trained by unsupervised NN to get their significant classes. Thirdly, the interesting rules are extracted by inspecting the attributes of customers. These rules are interesting and important for the traders, marketers and customers for making future business plan.

Index Terms —Heterogeneous data, Rule Extraction, Neural networks, Statistics.

I. INTRODUCTION
The Internet has provided a tremendous level of excitement through its involvement with all kinds of e-services such as e-commerce, e-business, e-supply, e-marketplace, e-payment, e-entertainment, e-learning, and so on. The e-commerce is now a hot and challenging topic. The growing interest in internet shopping encourages customers and organizations to know each other. Thus, organizations should understand their customer’s behavior, preferences and future needs. This leads many companies to develop a lot of e-service systems. In recent years, e-commerce has grown dramatically in terms of volume and variety of goods and services [1]. It is predicted by American International Telecommunication Union and American International Data Group that Internet trade will account for 42% of the global trade volume by 2010 [2].

An organization generates log files that track the steps of visitors as they move around on the shop websites. Analyzing these web log files is very challenging due to their huge size and the heterogeneity of the purchase patterns of each individual. Finding useful information buried in such large-scale datasets requires a rigorous analytical approach, yet the data are not very amenable to classical statistical models.

The discovery of interesting rules is to find the underlying patterns, correlation, associations and causal structures from the data [3]. Though it is very difficult to identify representative patterns of users’ behaviors on the Web, previous studies have attempted to develop rigorous analytical tools to investigate web visit patterns. Huberman [4] demonstrated that Web visits have some unique and regular patterns, particularly those related to the number of clicks (hyperlink requests) [5] explored a tool to describe such patterns in a visual way and defined statistical models for using in the data analysis. There is an another type of attempt that generates attributes considering various facts in the purchase patterns, such as length of pages, number of sub-pages, mode of purchase etc. [6].

There are a number of attempts to find out the inherent customer interest and/or relationship from their search navigation. Xiaobin et. al reported a system which actively monitors and tracks a user’s navigation [7]. Once a user’s navigation history is captured, they applied data mining techniques to discover the hidden knowledge contained in the history. The knowledge is then used to suggest potentially interesting web pages. A number of reports designed two measures of similarity and unexpectedness to analyze the
A. Examples of data

Here, we explain the visitors purchase patterns. In the first line of data, we see a series of numbers as follows. 591603924,(01,26),(01,26),(01,26),(01,34),(01,25),(46,17),(46,19),(46,19),(46,21),(46,18),
(46,20),(46,20),(46,22).


In this case, the web page of the shop is (01,26). Since commodities 21, 19 and 66 are on the same page, the data shows that he moved as such (01,26) (01,26) (01,26). The final page he reached was (46,22) and purchased the commodity. He visited a total number of 33 pages in order to purchase the commodity. This is how the customer searches the necessary pages to purchase the commodity. The purchase of commodities 21, 19 and 66 are on the same page, the data shows that he moved as such (01,26) (01,26) (01,26).

B. Features of data

The data are very heterogeneous as shown in TABLE I. The page address is a pair of integer values. A customer visits continuously from one page to another. It is also possible to come back to the previously visited pages. The transition of pages of a customer is not the same with another customer. But the number of pages a customer visits may be same with others. The percentage of the same number of pages is about 3. The number of pair varies from 4 to 649. First 20 customers with their number of attribute are listed in TABLE I.

C. Problems of data to analyze

The main problem of the customer’s purchase patterns is that they are raw and heterogeneous. The raw data means the unprocessed and unsuitable to use in a NN. This is quite impossible to use these kinds of data in neural network and to make a decision about the category of data or customers. We need to make equal size data for using in neural network. Not only that, we need a single attribute data – not a pair data, since neural network cannot use pair wise data. Moreover, first data point in a pair is more important than the second one. It is difficult to process such type of importance by a neural network.

III. METHODOLOGY

In this section the processes of extracting rules from heterogeneous data will be discussed. The entire process can be visualized from the Fig. 1. Several subsections of the methodology may be outlined as data preprocessing, clustering and rule extraction.

Firstly, data preprocessing step attempts to generate uniform data from raw data. It is very difficult to analyze these raw data. These raw data cannot be used directly in NN. So, it is necessary to create a standard data set that can be applied in NN. We will describe the way to generate equal length data from heterogeneous data by several steps as will be described in Sec. IV.

Secondly, since nothing is known about the category of data, one needs to group them. This step will cluster the entire customers into several groups. An unsupervised competitive neural learning algorithm is used to perform the clustering. Once the groups or clusters are available, it is easy to inspect the inherent relationship among the cluster and attributes.
Thirdly, the rules are extracted from the clusters. To analyze the relationship between customer profile and purchased products, LHS (conditional part) of rules consists of customer behavioral variables, and RHS (consequent part) consists of clusters.

The entire procedure is explained one by one from the next section.

IV. DATA PREPROCESSING

This section describes suitable data generation from raw data with several statistical justifications.

A. Joining pair wise data

We need to make a single attribute from a pair of attributes. But it is difficult to unite a simple arithmetic operation. Such as (02, 10) and (05, 07) are two different visited web pages. But by simple addition we get the same result (such 12). We just put a decimal point between two attributes. Therefore, they become new single attributes as 02.10 and 05.07. We combine pair wise data from heterogeneous data by adding a decimal point between them. We consider first part of the pair is more significant than the second part. Now, one can easily identify first part of the pair and second part also. So it is not unreasonable.

B. Data resizing

The number of data varied from 4 to 649. We take a standard data length for creating equal length all the data set. In this research, we have taken target length (TL) or desired length is 80. If the numbers of raw data are too less from desired length and the number of raw data are too high from desired length, it becomes difficult to make desired length. For this reason, we consider the number of data from 21 to 300. We filter the customers with attributes less than 21 and greater than 300. At last we consider 116 purchased patterns from 153 purchased patterns. Then rescale the filtering data set. These new resized data set (RDS) are performed in the next section.

C. Creating equal length data from heterogeneous length

We need to make equal length data for using in NN and CCA. We create equal length from the total heterogeneous data. It is the core work in data preprocessing. Now we will describe how to generate equal length data from different length data. Firstly, a RDS is taken. The original data length (DL) and TL are compared. If DL is equal to or less than TL, Attribute addition algorithm (AAA) is attempted, otherwise Attribute reduction algorithm (ARA) is performed. Here this method will be described briefly [11].

\[
\text{position index to copy a data} = \frac{\text{upper nearest ten}}{\text{number to reach TM} + 1} \ldots (1)
\]
Figure 2 shows the complete flow chart to create equal length data. AAA and ARA are explained in the following section. Store this data set in a file after performing AAA or ARA. Repeat the whole process for all filtering data set and store in a file. This file contains the training set for NN.

In AAA, at first we create TM data. That means, created data should be divisible by ten. A hypothetical example is explained in Fig. 4. Here raw data points are 8 and the nearest TM number is 10. From equation (1), we have position index as 10/(2+1) = 3.33 and it is greater than 3. Therefore, we consider data repeating position is 4 as shown in Fig. 4 to create additional two data. It is not unreasonable because of some reasons: a) we repeat some intermediate data so that there is no need to create new unknown attributes, b) we repeat data points for some intervals (not regular). So we can say that the data characteristics cannot be changed after performing TM. We create TM because to reduce computational complexity in the upcoming computation.

\[ f = f(x_0) + (x-x_0)f'[x_0,x_1] + (x-x_0)(x-x_1)f[x_0,x_1,x_2] + \cdots + (x-x_0)...(x-x_{n-1})/f[x_0,x_1,...x_n] \]

Figure 5 shows the sample of purchase pattern consisting of 30 attributes and Fig. 6 shows the generated 80 attributes using AAA from 30 attributes. The purchase pattern contain only 30 attributes, we have to generate additional 50 attributes using AAA. We see that Fig. 5.1 and Fig. 5.3 show similar shape. They are functionally approximately same.

ii) Attribute reduction algorithm (ARA)

ARA based on a comparison two neighboring data. In other word, it can be called comparison algorithm. A number of attributes are deleted comparing with a pre-specified threshold as shown in Fig. 7. Two neighboring attributes (say \( m \) & \( n \)) are compared. When the result does not exceed or equal the threshold value, attribute \( n \) is deleted, otherwise there is no deletion. The threshold value is not constant. Let’s have 1,2,3,3.25,4,5,6 data points and the threshold value is 0.5. Since 3~3.25 \( \leq 0.5 \), \( n=3.25 \) data point is deleted. This process will continue before the time when DL is equal to TL. If DL does not reach TL in beginning, this algorithm increases the threshold value of a step of 0.05 (or any suitable value).
ARA has performed approximately unbiased because of some reasons, a). We select zero or very low threshold value. For this reason, repeated attributes or similar attributes are deleted. b). ARA compares two neighboring attributes (say \( m \) & \( n \)). When attribute \( n \) is deleted, the next comparison takes a new set. For this reason, a series of repeated data cannot be deleted at a time. It performs all over the data set. c). If DL does not reach TL, this algorithm increases the threshold value very small. Relatively same attribute or same information is deleted but different information exists in the data set. So, we can say that ARA is not unreasonable.

Figure 8 shows two samples of purchase pattern consisting of 200 attributes and Fig. 9 shows the generated 80 attributes by ARA from 200 attributes. Fig. 8 and Fig. 9 show approximately similar transition. In this case, 120 attributes reduce by ARA to produce 80 attribute. In this way, the algorithm creates equal length purchase patterns with equal size of 80. One can easily understand by visual inspection that both (raw data and created data) sets that generated data set are equivalent.

![Figure 8: A sample of 200 attributes.](image1)

![Figure 9: Generated 80 attributes by ARA for 200](image2)

**D. Statistical validation**

We have developed an algorithm that can automatically generate equal length data form heterogeneous data. Without data preprocessing, one cannot use it in NN or any suitable networks. In order to apply it to NN, the number of data points or attributes of the data sets must be always same. In order to prove that the generated data points are approximately unbiased, we test original and generated data using statistical tests. This will prove that the generated data is statistically sound and functionally same with original heterogeneous data. The generated data are validated by three statistical tests.

i) The mean and standard deviation (SD): They are computed and listed in TABLE II & TABLE III for AAA and ARA respectively. Eight original data sets are chosen from different places of the entire data set. Four data points are used in AAA and other ones are used in ARA. The values of mean and SD for the original and generated data set are approximately same. For example, a data of length 30, the means are 24.0917 and 24.2103 for original and generated data sets respectively. Similarly, they have approximately same SDs. We take 95,139,200 and 227 attributes data set for ARA. For example, a data length of 200, the mean and SD for original data set are 25.8761 and 19.0904 respectively and generated data set are 25.6344 and 19.1331 respectively. This data has approximately same mean and SD. Similarly, we get good result for ARA. The result of data length 236 is quite different between original and generated data. Since a large data size of 227 is reduced to expected length of 80.

![TABLE II](image3)

<table>
<thead>
<tr>
<th>Original data point</th>
<th>Mean for original data</th>
<th>Mean for generated data</th>
<th>SD for original data</th>
<th>SD for generated data</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>24.0917</td>
<td>24.2103</td>
<td>22.2616</td>
<td>21.8562</td>
</tr>
<tr>
<td>42</td>
<td>32.2679</td>
<td>32.0436</td>
<td>17.7832</td>
<td>16.9584</td>
</tr>
<tr>
<td>57</td>
<td>17.5521</td>
<td>17.4035</td>
<td>21.2742</td>
<td>21.2265</td>
</tr>
<tr>
<td>76</td>
<td>18.9068</td>
<td>19.0723</td>
<td>19.0764</td>
<td>19.0242</td>
</tr>
</tbody>
</table>

![TABLE III](image4)

<table>
<thead>
<tr>
<th>Original data point</th>
<th>Mean for original data</th>
<th>Mean for generated data</th>
<th>SD for original data</th>
<th>SD for generated data</th>
</tr>
</thead>
<tbody>
<tr>
<td>95</td>
<td>30.3716</td>
<td>30.7741</td>
<td>18.1078</td>
<td>18.1856</td>
</tr>
<tr>
<td>139</td>
<td>18.6832</td>
<td>18.2469</td>
<td>19.9058</td>
<td>19.3794</td>
</tr>
<tr>
<td>200</td>
<td>25.8761</td>
<td>25.6344</td>
<td>19.0904</td>
<td>19.1331</td>
</tr>
<tr>
<td>227</td>
<td>17.2673</td>
<td>16.5151</td>
<td>22.7767</td>
<td>23.1510</td>
</tr>
</tbody>
</table>

ii) Z-test: A Z-test is a statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution. Due to the central limit theorem, many test statistics are approximately normally distributed for large samples. Therefore, many statistical tests can be performed as approximate Z-tests if the sample size is not too small.

The most general way to obtain a Z-test is to define a numerical test statistic that can be calculated from a collection of data, such that the sampling distribution of the statistic is approximately normal under the null hypothesis. Statistics that are averages (or approximate averages) of approximately independent data values are generally well-approximated by a normal distribution. An example of a statistic that would not be well-approximated by a normal distribution would be an extreme value such as the sample maximum [13].
If $\bar{x}_1$ and $\bar{x}_2$ are the means of two independent random samples, then the sampling distribution of the statistic $\bar{x}_1 - \bar{x}_2$ has the mean $\mu_1 - \mu_2$ and the standard deviation.

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Where $\mu_1, \mu_2, \sigma_1$, and $\sigma_2$ are the mean and SD of the two sampled. If limit to large samples, $n_1 > 30$ and $n_2 > 30$, we can base the test of the null hypothesis that the same mean on the statistic. It has approximately the SD normal distribution.

We need to find the value of $Z$ that will only be exceeded 5% of the time since we have set our alpha level at 0.05. Since the Z score is normally distributed (or has the Z distribution). For example, to get the $Z$ for a 95% confidence interval, make the shaded area 0.95. Then choose “between” and you will see that 0.95 of the area is between -1.96 and 1.96. For null hypothesis, it is seen that the Z value exceed this value or not. In TABLE IV, first column shows the raw data, second column shows the degree of freedom, third column shows the Z-test value and finally we show that it is accepted or rejected. Here two data sets are taken for testing at a time, one is original raw data and second one is corresponding generated data. The size of the generated data is 80 and it is fixed throughout the calculation. The tests exhibit that the original and generated data are similar.

### Table IV: Statistical Validation with Z-test

<table>
<thead>
<tr>
<th>Number of raw data</th>
<th>Degree of freedom</th>
<th>Z-test values</th>
<th>Test hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>108</td>
<td>0.0252</td>
<td>Accepted</td>
</tr>
<tr>
<td>57</td>
<td>135</td>
<td>0.0404</td>
<td>Accepted</td>
</tr>
<tr>
<td>42</td>
<td>120</td>
<td>0.0683</td>
<td>Accepted</td>
</tr>
<tr>
<td>76</td>
<td>154</td>
<td>0.0542</td>
<td>Accepted</td>
</tr>
<tr>
<td>95</td>
<td>173</td>
<td>0.1462</td>
<td>Accepted</td>
</tr>
<tr>
<td>139</td>
<td>217</td>
<td>0.1565</td>
<td>Accepted</td>
</tr>
<tr>
<td>200</td>
<td>278</td>
<td>0.0956</td>
<td>Accepted</td>
</tr>
<tr>
<td>227</td>
<td>305</td>
<td>0.1686</td>
<td>Accepted</td>
</tr>
</tbody>
</table>

### Table V: Polynomial Curve Fitting Result for AAA

<table>
<thead>
<tr>
<th>Data points</th>
<th>Coefficients of polynomial fitting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a_4$</td>
</tr>
<tr>
<td>Ori. (30) data</td>
<td>-0.0099</td>
</tr>
<tr>
<td>Gen. (80) data</td>
<td>-0.0000</td>
</tr>
<tr>
<td>Ori. (57) data</td>
<td>-0.0001</td>
</tr>
<tr>
<td>Gen. (80) data</td>
<td>-0.0000</td>
</tr>
</tbody>
</table>

### Table VI: Polynomial Curve Fitting Result for ARA

<table>
<thead>
<tr>
<th>Data</th>
<th>Coefficients of polynomial fitting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a_4$</td>
</tr>
<tr>
<td>Ori. (95) data</td>
<td>-0.00</td>
</tr>
<tr>
<td>Gen. (80) data</td>
<td>-0.00</td>
</tr>
<tr>
<td>Ori. (200) data</td>
<td>0.00</td>
</tr>
<tr>
<td>Gen. (80) data</td>
<td>0.00</td>
</tr>
</tbody>
</table>

### V. Clustering and Rule Extraction

Rule induction is an area of machine learning in which formal rules are extracted from a set of observations. The rules extracted may represent a full scientific model of the data, or merely represent local patterns in the data. Some rule induction paradigms are: Association rule algorithms, Decision rule algorithms, Hypothesis testing algorithms, Horn clause induction, Version spaces, Rough set rules, Inductive Logic Programming, and so on. We use decision rule algorithm from these algorithms.

#### A. Clustering with neural network

After preprocessing the data set it is usable for cluster analysis. Clustering is the classification of objects into groups (called clusters) so that objects from the same cluster are more similar to each other than objects from different clusters. Clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics [15]. Neural networks are widely used for clustering data. An unsupervised competitive learning network (CLN) is used for clustering all customers.

After training these homogeneous data through the CLN a significant cluster are found as shown in Fig. 10. From this output six significant groups are extracted. The output neuron 4 wins the maximum number of customers. So this cluster contains the general trends of customers. On the other hand, the neuron 5 wins the minimum number of customers.
Figure 10: Output of CLN after training.

TABLE VII
CHARACTERISTICS OF ESTIMATED CLUSTERS

<table>
<thead>
<tr>
<th>Cluster name</th>
<th>Winning neuron</th>
<th>Total customers</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>36</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
<td>16</td>
</tr>
<tr>
<td>D</td>
<td>7</td>
<td>18</td>
</tr>
<tr>
<td>E</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>F</td>
<td>5</td>
<td>8</td>
</tr>
</tbody>
</table>

TABLE VII shows the summary of clusters. The maximum number of customers contributes in cluster B. Only four purchased patterns (one at output neuron 9 and three at 10) cannot occupy in any group (as of Fig. 10). We ignored them for simplicity sake.

B. Rules

A set of decision rules is the verbal equivalent of a graphical decision tree, which specifies class membership based on a hierarchical sequence of (contingent) decisions. Each rule in a set of decision rules therefore generally takes the form of a Horn clause wherein class membership is implied by a conjunction of contingent observations. A rule can be defined as

\[ \text{IF condition1 AND condition2 AND/ OR ... AND condition m THEN CLASS = class} \]

Where condition is in general contingent on the choice of condition1. Decision rules can be transcribed from the corresponding decision tree, or can be induced directly from observations. Decision rules are commonly used in the medical field. For example, the Ottawa Ankle Rules guide obtaining radiographs for traumatic ankle pain [16].

Interesting rules are listed in Fig. 11 & 12. We find out these rules among the distinct customer, which are representing the total characteristic of data. Here it is seen that there are two major rules - Clustering rule and General rule. They are described below.

i) Clustering rule

The clustering rules are defined as those rules which are directly related with individual cluster. For example, if the data points \((31.05 \text{ or } 1.26 \text{ or } 3.25) \text{ and } (1.34 \text{ or } 1.72 \text{ or } 3.80 \text{ or } 3.23) \text{ and } 46.17 \text{ and } 46.19 \text{ and } 45.02 \text{ and/or } 46.21 \text{ and } 46.18 \text{ and } 46.20 \text{ and } 46.22)\) then cluster A.

Rule 1:
\[ \text{IF } (31.05 \text{ or } 1.26 \text{ or } 3.25) \text{ and } (1.34 \text{ or } 1.72 \text{ or } 3.80 \text{ or } 3.23) \text{ and } 46.17 \text{ and } 46.19 \text{ and } 45.02 \text{ and/or } 46.21 \text{ and } 46.18 \text{ and } 46.20 \text{ and } 46.22 \text{ THEN } \text{CLUSTER = A} \]

Rule 2:
\[ \text{IF } ((31.50 \text{ or } 0.54) \text{ and } 46.17 \text{ and } (44.97 \text{ or } 44.91 \text{ or } 44.79) \text{ and } 46.19 \text{ and } 45.02 \text{ and/or } 46.21 \text{ and } 46.20 \text{ and } 46.22) \text{ THEN CLUSTER = B} \]

Rule 3:
\[ \text{IF } ((31.5 \text{ or } 1.51 \text{ or } 2.97) \text{ and } (4.05 \text{ or } 4.63 \text{ or } 3.8) \text{ and } (39.94 \text{ or } 44.91) \text{ and } 46.17 \text{ and } 45.02 \text{ and } 46.19 \text{ and } 46.18 \text{ and/or } 46.20 \text{ and } 46.22) \text{ THEN CLUSTER = C} \]

Rule 4:
\[ \text{IF } (31.50 \text{ and } 39.94 \text{ and } 46.17 \text{ and } 45.02 \text{ and } 46.04 \text{ or } 3.25 \text{ or } 0.36) \text{ and } 46.19 \text{ and } 46.18 \text{ and/or } 46.20 \text{ and } 46.22) \text{ THEN CLUSTER = D} \]

Rule 5:
\[ \text{IF } ((31.5 \text{ or } 3.2) \text{ and } (3.74 \text{ or } 3.25 \text{ or } 4.20) \text{ and } (1.3 \text{ or } 4.61 \text{ or } 4.7 \text{ or } 4.5 \text{ or } 3.25) \text{ and } 46.17 \text{ and } 46.19 \text{ and } 45.02 \text{ and } 46.18 \text{ and } 46.20 \text{ and } 46.22) \text{ THEN CLUSTER = E} \]

Rule 6:
\[ \text{IF } (31.5 \text{ or } 2.79) \text{ and } (39.94 \text{ or } 40.64) \text{ and } (1.7 \text{ or } 44.91 \text{ or } 42.06) \text{ and } 45.02 \text{ and } 46.17 \text{ and } 45.19 \text{ and } 46.18 \text{ and } 46.20 \text{ and } 46.22) \text{ THEN CLUSTER = F} \]

Fig. 11: Interesting clustering rules

ii) General rule

The general rules are the ones which have been observed in several clusters together. The rules 7, 8 and 9 of Fig. 12 are contributing the general rule. If the data points follow the following sequence it contributes the all clusters.

\[ \text{IF } (31.50 \text{ and } 46.17 \text{ and } 45.02 \text{ and } 46.19 \text{ and } 46.18 \text{ and/or } 46.20 \text{ and } 46.22) \text{ THEN CLUSTER = A} \]

\[ \text{IF } (31.50 \text{ and } 46.17 \text{ and } 45.02 \text{ and } 46.19 \text{ and } 46.18 \text{ and/or } 46.20 \text{ and } 46.22) \text{ THEN CLUSTER = B} \]

\[ \text{IF } (31.50 \text{ and } 46.17 \text{ and } 45.02 \text{ and } 46.19 \text{ and } 46.18 \text{ and/or } 46.20 \text{ and } 46.22) \text{ THEN CLUSTER = C} \]

\[ \text{IF } (31.50 \text{ and } 46.17 \text{ and } 45.02 \text{ and } 46.18 \text{ and/or } 46.20 \text{ and } 46.22) \text{ THEN CLUSTER = D} \]

\[ \text{IF } (31.50 \text{ and } 46.17 \text{ and } 45.02 \text{ and } 46.19 \text{ and } 46.18 \text{ and/or } 46.20 \text{ and } 46.22) \text{ THEN CLUSTER = E} \]

\[ \text{IF } (31.50 \text{ and } 46.17 \text{ and } 45.02 \text{ and } 46.18 \text{ and/or } 46.20 \text{ and } 46.22) \text{ THEN CLUSTER = F} \]

The meaning of \(A(41, 93)\) is that the antecedents of the rule are mostly found in customer 41 and 93.
Here we show the clusters with their IDs. In the same way we will explain the rule 7 and 8 of Fig. 12. It is seen that the most frequently searched web pages are 31.50, 3.25, 46.17, 46.19, 45.02, 46.18, 46.20, and 46.22. After getting these rules, it is necessary to determine which page number is most important or which products are demandable for customers. Now traders can determine which rule should be the general sequence for all clusters. If trader can predict the customer demand, he can understand which product is more necessary or less necessary to store.

C. Pruning of rules
The rules obtained in Figs 11 & 12 are in many senses long and spurious. They may contain unnecessary web pages or items which in fact are redundant. That’s why we can reduce the size of the rule by inspecting a number of points such as i) items are most frequently visited pages or not, ii) Any one item or more of OR part of the rules may be unimportant. On the basis of this we have pruned them and listed in Fig. 13 as concise ones. Finally we have six rules in total. We have pruned the rule 3, 5, and 6 of Fig. 11 and rule 1, 2, and 4 are taken. These three rules are arranged as rule 1, 2, and 3 in Fig 13. The reason for removal of rule 3, 5 and 6 is that there are less number of customers in those clusters such as in C, E and F (see Table VII).

Moreover, some attributes of rules are ignored due to their less frequent appearance in the search histories. The attributes 1.26 and 3.80 of Rule 1 of Fig 11 are removed to form Rule 1 of Fig. 13. Attribute 44.91 of Rule 2 of Fig. 11 is removed to form Rule 2 in Fig. 13. The attribute 0.36 is removed of Rule 4 of Fig. 11 to form Rule 3 of Fig. 13. Rules 4, 5, and 6 of Fig. 13 are separated from 7, 8, and 9 of Fig. 12. Several ‘OR’ components and attributes are dropped to form the final rules.

Similarly, the final three rules 4, 5, and 6 of Fig 13 are extracted from rule 7, 8 and 9 of Fig. 12. It is seen that there are six rules in Fig. 12 for rule 9. However, we can induce two from them as depicted Rule 6 in Fig. 13. Here, only rules for cluster B and D are taken. This is because these two clusters are dominant by the number of customers and maximum number of customers in the rules. Similarly, Rule 5 of Fig. 13 is inferred from Rule 8 of Fig. 12. For the same reason we choose Rule 4 of Fig. 13 as a dominant one from Rule 7 of Fig. 12.

D. Determination of page significance
One of the goals of rule extraction is to determine the important attributes from the extracted rules. One can easily determine the AND OR logical flow as shown in Fig. 14. Taking one from OR part and all from AND switch is enough for the production of a significant page. One can consider this as a quality of a rule. The AND part is very important in order to fulfill the convergence of a rule. By visual inspection and induction the significant pages are 3.25, 31.50, 45.02, 46.17, 46.18, 46.19, 46.20, and 46.22. These pages are important because they are seen in almost every rule. What does the actual page contain is in fact immaterial for the current interest. We just analyze the attributes of the pages.

Rule 7:
If (31.50 and 46.17 and 46.19 and 45.02 and 46.21 and 46.18 and 46.20 and 46.22) then cluster D (15, 78), OR
If (31.50 and 46.17 and 46.19 and 45.02 and 46.21 and 46.18 and 46.20 and 46.22) then cluster E (21, 30, 55)

Rule 8:
If (3.25 or 0.54 or 1.26 or 2.79 and 46.17 and 45.02 and 46.19 and/or 46.21 and 46.18 and 46.20 and 46.22) then cluster A (1, 24, 100), OR
If (3.25 or 0.54 or 1.26 or 2.79 and 46.17 and 45.02 and 46.19 and/or 46.21 and 46.18 and 46.20 and 46.22) then cluster B (57)

Rule 9:
If (31.50 and 46.17 and 46.19 and 45.02 and 46.18 and/or 46.20 and 46.22) then cluster A (41, 93), OR
If (31.50 and 46.17 and 46.19 and 45.02 and 46.18 and/or 46.20 and 46.22) then cluster B (20, 94, 107), OR
If (31.50 and 46.17 and 46.19 and 45.02 and 46.18 and/or 46.20 and 46.22) then cluster C (42, 48, 101) OR
If (31.50 and 46.17 and 46.19 and 45.02 and 46.18 and/or 46.20 and 46.22) then cluster D (15, 60, 78, 81) OR
If (31.50 and 46.17 and 46.19 and 45.02 and 46.18 and/or 46.20 and 46.22) then cluster E (21, 30, 33, 55) OR
If (31.50 and 46.17 and 46.19 and 45.02 and 46.18 and/or 46.20 and 46.22) then cluster F (67, 79)

Rule 1:
If (((31.05 or 3.25) and (1.34 or 1.72) and 46.17 and 46.19 and 45.02 and/or 46.21 and 46.18 and 46.20 and 46.22) then cluster A)

Rule 2:
If (31.50 and 46.17 and (44.97 or 44.79) and 46.19 and 45.02 and 46.18 and 46.20 and 46.22) then cluster B

Rule 3:
If (31.05 and 39.94 and 46.17 and 45.02 and (4.61 or 3.25) and 46.19 and 46.18 and/or 46.20 and 46.22) then cluster D

Rule 4:
If (31.50 and 46.17 and 46.19 and 45.02 and 46.21 and 46.18 and 46.20 and 46.22) then cluster E (21, 30, 55)

Rule 5:
If (3.25 or 1.26 and 46.17 and 45.02 and 46.19 and/or 46.21 and 46.18 and 46.20 and 46.22) then cluster A (1, 24, 100), OR

Rule 6:
If (31.50 and 46.17 and 46.19 and 45.02 and 46.18 and/or 46.20 and 46.22) then cluster D (15, 60, 78, 81)

Fig. 12: Interesting general rules

Fig. 13: Filtered rules
Fig. 14: AND OR switch for significance determination

The number of attributes of finally obtained rules varies from 6 – 9. Because NN clustering was used the rules belonging to the a cluster have higher correlation than others. However, finally the traders/enterprises can choose frequently visited attributes/web pages.

VI. COMPARATIVE STUDIES

The process of discovering interesting and unexpected rules from data set is known as association rule mining. The technique was first introduced by Agrawal [17]. There are vigorous number of attempts to discover these rules. We compare the proposed technique with several of them. The mining change of customer (MCC) [8] and a similar approach mining change in market (MCM) [9] have discussed the issue of basically transition of customer choices. Another approach describes the online monitoring the interesting customer using clustering, called ‘SurfLen’ [7]. The main difference between this SurfLen and others is that SurfLen attempts on online monitoring, while others do not. The association rules are discovered with genetic algorithm (RGA) in [18]. An overview of their strategies including the proposed technique can be visualized from TABLE VIII. It is very difficult to compare them one to one. Because the set ups and focuses of different algorithms are different. For example, ‘SurfLen’ monitors customer choice, while RGA is finds frequently visited pages. The proposed technique works on the data taken from internet shopping and attempts to analyze them. A limitation of the proposed technique is its difficulty in processing a transaction with very small number of attributes or very large number of attributes, relative to an average number.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Rule type</th>
<th>Technical base</th>
<th>Real time</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCC</td>
<td>Association</td>
<td>Similarity</td>
<td>Offline</td>
</tr>
<tr>
<td>MCM</td>
<td>Association</td>
<td>Similarity</td>
<td>Offline</td>
</tr>
<tr>
<td>SurfLen</td>
<td>Association</td>
<td>NN cluster</td>
<td>Online</td>
</tr>
<tr>
<td>RGA</td>
<td>Association</td>
<td>GA</td>
<td>Offline</td>
</tr>
<tr>
<td>Proposed</td>
<td>Association</td>
<td>NN cluster</td>
<td>Offline</td>
</tr>
</tbody>
</table>

VII. CONCLUSION

This paper presents a framework that extracts interesting rules from internet search histories of customers. The Internet search histories are heterogeneous in length. So firstly, same length patterns from heterogeneous patterns are created and they are generated with two ways - attribute addition and attribute reduction. We applied the generated equal length data in an unsupervised neural network in order to find the possible clusters. From the clusters we have found out interesting rules by inspecting the attributes of each customer. The rules are important especially for the online traders.

Online traders often want the interesting web sites which are attractive to their customers. The important search patterns of customers from their search histories are sorted out. The important rules and attributes are found out in this framework. The most important web pages from the investigations of induced rules are 31.50, 3.25, 46.17, 46.19, 45.02, 46.18, 46.20, and 46.22. This will help to find the link between traders or web pages. The results are important for the policy makers of companies or organizations to adapt online customer’s interest.

ACKNOWLEDGMENT

The authors wish to thank anonymous reviewers for their constructive comments which helped a lot to improve the paper.

REFERENCES


