A Methodology for Extracting Head Contents from Meaningful Tables in Web Pages

Madhuri M. Chavan
M.E. student
D. Y. Patil College of Engg. & Tech.,
Kolhapur, Maharashtra, India.
Email: replymadhu@sify.com

S. K. Shirgave
Asst. Prof. Dept. of I.T
D. K. T. E. College, Ichalkaranji.
Maharashtra, India.
Email: skshirgave@yahoo.com

Abstract
Tables are an important feature of presenting information & are widely used on the web. They show relational data in a simple & precise manner. A typical web page consists of many blocks or areas e.g. main content areas, advertisements, images etc. Tables contain meaningful information. Almost all data is arranged in tabular format. Tables describe relational information in a compact manner. So there is need to find out the tables which contains meaningfulness structural data. In this paper, a method is introduced for determining the meaningfulness of a table and extracting the Head from meaningful table.

Keywords: DOM Tree, Text mining, Table mining, information extraction, Web table

I. INTRODUCTION
Today, the rapid expansion of the internet has made a www a popular resource for collecting information. Web contains huge amount of web pages. Extracting the information which is spread over the internet is a part of web information extraction. From this large amount of information users wish to extract the specific information. And to get that information users has to process the data present on web pages. Text mining is one of the tasks which is based on the information extraction rules, locates the specific information. Text mining is carried out by using tags and other features, and HTML documents are made by different tags. Thus, HTML documents on the internet gives an idea in identifying the rules required for text mining. Tables are interesting because they present information in simple & in well-structured manner. Table contains useful information hence they are frequently used in web documents. These tables are known as web tables categorized as meaningful tables & decorative tables. Tables which contain relational data means the data present in attribute-value pair, these tables are known as meaningful tables. Structural characteristics of HTML documents distinguish the meaningful tables from decorative tables. 2-D tables are considered as meaningful tables and 1-D tables can be considered as decorative tables. Table HEAD abstracts the data, hence extracting meaningful information from table include a task which extracts the HEAD components from table. Table filtering, table recognition, table interpretation these are the key terms required for table processing. Different rules & characteristics are needed to consider in determining the meaningful tables. The system model present here considered these rules & separates the meaningful tables from decorative tables and then extracts the table HEAD.

II. RELATED WORK
Table mining has become a widely researched topic on its own over the years. Table information extraction is a sub domain of the information extraction process. Research into Web table mining can be classified into domain-specific research and domain-independent research. Domain-independent approaches have recently been introduced into table mining. A previous researcher, Chen et al.[4] considered the term “table mining” for table information extraction. They employ heuristic rules to filter out non-genuine tables from their test set and make assumptions about cell content similarity for table recognition and interpretation. Tengli et al. [10] present an algorithm that extracts tables and differentiates between label and data cells. Yalin Wang and Hu [7] train a classifier on content features of individual cells and non-text layout features from the HTML source to perform the same task of table location. They have attempted to implement a general table mining system using a machine- learning algorithm. They applied information retrieval (IR) strategies to their table mining. However, this strategy could not cope with new tables that contained unknown words. Several studies [4], [6] have extracted table information using extraction rules according to a special tabular form. Because these studies dealt only with such forms, the researchers experienced difficulties in coping with the various Web document formats.

Sung-Won Jung and Hyuk-Chul Kwon. [1] have established a preprocessing method for determining the meaningfulness of a table to allow for information extraction from tables on the Internet. However, tables are used on the Internet for both knowledge structuring and document design. Therefore, it becomes important task to determine whether or not a table has meaningfulness that is related to the structural information provided at the level of the table head. Accordingly, they have investigated the types of tables present in HTML documents, established the features that distinguished meaningful tables from others, constructed a training data set using the established features after having filtered any obvious decorative tables.
and constructed a classification model using a decision tree. They have considered the appearance features and consistency features for identifying types of the table. But their attempt could not cope with a table that does not contain head and they are failed to extract an appropriate HEAD using background color and font. The filtering rules and patterns used by both the researchers require updates. In paper [1], the filtering rules are used which are limited and the system model are having following limitations:

a) Extracting an appropriate HEAD using background color.
b) Incorrect head extraction i.e. in the absence of HEAD component.
c) Extracting an appropriate HEAD using background color and font.

Due to the incorrect filtering rules proper distinguishing of meaningful tables from decorative tables is not carried out.

The objective of this work is to apply table mining to general HTML documents, separates meaningful tables from decorative tables, extract the information using HEAD and to overcome the limitations of the existing system[1]. For that purpose we applied a structured and modified filtering rules to the system model.

III. DEFINITION AND CHARACTERISTICS OF WEB TABLES

Web table:
In web pages, data is arranged in tabular format that structure is called as web tables. The tags <table> and </table> represents the starting and ending tags of the tables respectively.

Meaningful table:
The data which represents in attribute-value pair, that data is called as relational data. Tables contains such relational data are known as meaningful table. Multiple columns and multiple rows is the structural characteristic of the meaningful table. These tables contain numerical, text data.

Decorative table:
One-dimensional tables are considered as decorative tables. Decorative tables contain other tables as content. These tables normally contain images, links.

IV. PROPOSED WORK

The techniques prescribed in the previous work are used to extract head components from web tables. But in some cases they are failed in extracting head components. So for proper head extraction the proposed model is designed. The proposed system model is shown in fig.

![Architecture of Proposed System Model](image)

The Proposed system model includes following modules:

1. Page collection process:

   Pure HTML pages are selected by removing extra scripting. Web pages contain data such as hyperlinks, images, script. So it is necessary to remove such unwanted script if any, during the time when a page is selected for processing. The HTML code is then transferred into XML code in order to generate DOM Tree.

   DOM tree representation: Web information extraction commonly focuses on detecting certain information patterns in the Dom tree of web pages.

![DOM Tree](image)
In this process the XML code is then converted into a DOM Tree, where each element is separated, based on Document, Root element, Element, Attribute and Text.

2. **Transformation Process :**

(a) Extraction of meaningful tables:

This process makes use of the classifier called as Decision Tree [11]. After the filtration step the data is passed to the decision tree classifier. Definition of decision tree is as follows:-

**Definition:** Given a database \( D = \{t_1, t_2, \ldots, t_n\} \) where \( t_i = \langle c_{i1}, c_{i2}, \ldots, c_{in} \rangle \) and the database schema contains the following attributes \( \{A_1, A_2, \ldots, A_h\} \). Also given is a set of classes \( C = \{C_1, C_2, \ldots, C_m\} \). A decision tree (DT) or classification tree is a tree associated with \( D \) that has the following properties:

1. Each internal node is labeled with an attribute, \( A_i \)
2. Each arc is labeled with a predicate that can be applied to the attribute associated with the parent.
3. Each leaf node is labeled with a class, \( C_j \)

The Process of Constructing a Decision Tree

- Select an attribute to place at the root of the decision tree and make one branch for every possible value.
- Repeat the process recursively for each branch.

**Decision tree algorithm is as follows:**

**Algorithm:-**

Input: \( D \)  // Training data set
Output: \( T \)  // Decision tree

**DTBuild algorithm:**

1. If \( T = \text{null} \), create root node and label with splitting attribute;
2. If \( T \) is an internal node, add arc to root node for each split predicate and label;
3. For each arc do:
   - Create leaf node and label with appropriate class;
   - If stopping point reached for this path, then
      - Create leaf node and label with appropriate class;
   - Else
      - \( T' = \text{DTBuild} \) (\( D \));
      - \( T = \text{Add} \ T' \) to arc

The ID3 technique to building a decision tree is based on information and attempts to minimize the expected number of comparisons. The concept used to quantify information is called entropy. Entropy is used to measure the amount of uncertainty in a data set. Certainly when all data in a set belong to a single class, there is no uncertainty. In this case the entropy is zero.

**Definition of entropy:**

Given probabilities \( P_1, P_2, P_3, \ldots, P_s \), where \( P_1, P_2, \ldots, P_s \) is entropy is defined as,

\[
H(F_1, F_2, \ldots, F_s) = \sum_{i=1}^{s} (P_i \log (1/P_i))
\]
Given a database state, \( D \), \( H(D) \) finds the amount of order in that state. When that state is split into new states \( S = \{ D_1, D_2, D_s \} \) again we check the entropy. The entropies of the split datasets are weighted by the fraction of the dataset being placed in that division. The ID3 algorithm calculates the gain of a particular split by the following formula.

\[
\text{Gain}(D,S) = H(D) - \sum_{i=1}^{n} \frac{P(D_i)H(D_i)}{P(D)}
\]

The decision tree algorithm C4.5 improves ID3 in following ways:

a) **Missing Data:** When the decision tree is built, missing data are simply ignored. That is the gain ratio is calculated by looking at the other records that have a value for that attribute.

b) **Splitting:** In the extreme case, an attribute that has a unique value for each tuple in the training set would be only one tuple for each division. An improvement can be made by taking into account the cardinality of each division. This approach uses the Gain Ratio as opposed to Gain.

The Gain Ratio is defined as,

\[
\text{Gain Ratio}(D,S) = \frac{H(D_j) - \sum\frac{|D_i|}{|D|}H(D_i)}{\sum\frac{|D_i|}{|D|}H(D_i) - \frac{|D|}{|D|}H(D)}
\]

For splitting purposes, C4.5 uses the largest Gain Ratio that ensures a larger than average information gain. Hence the largest value of the Gain Ratio decides the type of class. Above decision tree algorithm is used to separates the meaningful tables from decorative tables. Calculation of the gain ratio decides the meaningfulness of the table. Consider following example of HTML document

**Meaningful Table Generation:** After applying filtering rules the obvious decorative tables are filtered out from the web page and remaining data is then send to the C4.5 decision tree algorithm. Wherein the gain ratio is calculated. This gain ratio value will decide which tables are meaningful and which decorative tables are. It is observed that the gain ratio value is closer to the threshold value which becomes useful in distinguishing meaningful tables from decorative tables. Further, it is possible to include different content types like appearance features and consistency features.

**Appearance Features:**
- \(<\text{Th flag}>, <\text{caption flag}>, <\text{TablesInTable}>, <\text{img}>, <\text{a href}>, <\text{input}>, <\text{Frac_Text}>, <\text{Table Shape}>, \text{empty cells} \>

**Consistency Features:**
- Consistency features can be used to check the structure of the web tables i.e. the number of cells and their characteristics. Further these features will be useful in determining whether the table is decorative or meaningful table. If anyone of the feature is closer to zero, then it is considered to be meaningful table. Following list represents these features.

1. Standard deviation of number of columns
\[
dC = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (c_i - \bar{c})^2}
\]

2. Standard deviation of number of columns
\[
dR = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (r_i - \bar{r})^2}
\]

3. Standard deviation of cell length of the rows.
\[
dCLR = \frac{1}{\bar{r}} \sum_{i=1}^{n} \left( \frac{1}{\bar{c}_i} \sum_{j=1}^{\bar{c}_i} (c_{ij} - \bar{c})^2 \right)
\]

4. Standard deviation of cell length of the columns.
\[
dCLC = \frac{1}{\bar{c}} \sum_{i=1}^{n} \left( \frac{1}{\bar{r}_i} \sum_{j=1}^{\bar{r}_i} (r_{ij} - \bar{r})^2 \right)
\]

5. Consistency of content instance types of the rows.
\[
\text{CCR} = \frac{1}{\bar{r}} \sum_{i=1}^{n} \left( \sum_{j=1}^{\bar{r}_i} \frac{1}{\bar{c}_i} \sum_{k=1}^{\bar{c}_i} \text{diff}(cc_{ij}, cc_{ik}) \right)
\]

6. Consistency of content instance types of the columns.
\[
\text{CCC} = \frac{1}{\bar{c}} \sum_{i=1}^{n} \left( \sum_{j=1}^{\bar{c}_i} \frac{1}{\bar{r}_i} \sum_{k=1}^{\bar{r}_i} \text{diff}(ct_{ij}, ct_{ik}) \right)
\]

7. Tag consistency of the rows.
\[
\text{TCR} = \frac{1}{\bar{r}} \sum_{i=1}^{n} \left( \sum_{j=1}^{\bar{r}_i} \text{diff}(ct_{ij}, ct_{ij}) \right)
\]

Where,
• $c$ is the average number of columns for $\frac{1}{2} \sum_{i=1}^{n} c_i$.

• $c_i$ is the number of cells in rows $i$.

• $r_i$ is the number of rows.

• $r = \text{the average number of rows for } \frac{1}{2} \sum_{j=1}^{m} r_j$.

• $r_j$ is the number of rows in column $j$.

• $c$ is the number of columns.

• $c_{ij}$ is the number of characters in row $i$, column $j$.

• $c_i$ is the number of characters in row $i$.

• $c_{j}$ is the content-type of a cell in row $i$, column $j$.

• $c_{ij}$ is the average number of characters in column $j$.

• $c_{ij}$ is the tag-type of a cell in row $i$, column $j$.

• Diff ($c_1, c_2$) = 0 if $c_1 = c_2$ and $= 1$ otherwise.

After applying these features on the web page mentioned in shown in figure 3, the meaningful tables are extracted shown in figure 5.

<table>
<thead>
<tr>
<th>Sr No</th>
<th>Date of Advertisement</th>
<th>Details / Posts</th>
<th>Related Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>December 28, 2008</td>
<td>Abbrev. No. Est. (01) (11/04) 2008</td>
<td>General Conditions and Terms of Use Form</td>
</tr>
<tr>
<td>2</td>
<td>December 28, 2008</td>
<td>Abbrev. No. Ext. (01) (11/04) 2008</td>
<td>General Conditions and Terms of Use Form</td>
</tr>
<tr>
<td>3</td>
<td>December 30, 2008</td>
<td>SL No.</td>
<td>Approvals</td>
</tr>
<tr>
<td>5</td>
<td>27/05/2010</td>
<td>Abbrev. No. Ext. (01) (22/02/2010)</td>
<td>General Terms and Conditions - System Administrator</td>
</tr>
<tr>
<td>6</td>
<td>18/06/2011</td>
<td>Abbrev. No. Ext. (01) (22/02/2010)</td>
<td>General Terms and Conditions</td>
</tr>
<tr>
<td>7</td>
<td>30/05/2011</td>
<td>Abbrev. No. Ext. (01) (22/02/2010)</td>
<td>General Terms and Conditions</td>
</tr>
<tr>
<td>8</td>
<td>06/07/2011</td>
<td>Abbrev. No. Ext. (01) (22/02/2010)</td>
<td>General Terms and Conditions</td>
</tr>
</tbody>
</table>

Figure 5 Example of Meaningful table.

Decorative Table Generation:

<table>
<thead>
<tr>
<th>Appearance Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>CaptionText</td>
</tr>
<tr>
<td>TableHeader</td>
</tr>
<tr>
<td>Table Footer</td>
</tr>
<tr>
<td>Table Row</td>
</tr>
<tr>
<td>Table Data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Consistency Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
</tr>
<tr>
<td>ID</td>
</tr>
<tr>
<td>Type</td>
</tr>
<tr>
<td>Format</td>
</tr>
<tr>
<td>Language</td>
</tr>
<tr>
<td>Source</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Styling Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background Color</td>
</tr>
<tr>
<td>Border Color</td>
</tr>
<tr>
<td>Border Style</td>
</tr>
<tr>
<td>Text Color</td>
</tr>
<tr>
<td>Text Font</td>
</tr>
<tr>
<td>Text Size</td>
</tr>
<tr>
<td>Text Align</td>
</tr>
<tr>
<td>Text Weight</td>
</tr>
<tr>
<td>Text Style</td>
</tr>
<tr>
<td>Text Decoration</td>
</tr>
</tbody>
</table>

Figure 6 Example of Decorative table.

Figure 7 Snapshot of decision tree of meaningful table which is shown in figure 5.
Figure 7 shows the detailed information about all appearance and consistency features.

(b) Extraction of HEAD:-

In HTML, the tags <th>, <i>, <u>, <td bgcolor>, <font size>, <font face>, <font color> and their attributes offer clues in identifying the HEAD. The rows and columns containing the “rowspan” or the “colspan” attributes of the <td> tag can be a part of the HEAD. Hence the cells below the first row can also be considered as a part of the HEAD. <bgcolor> and <font size>, <font face> these tags are put together and are considered in head extraction i.e. the same priority is assigned to all these 3 tags. Hence if a web table contains HEAD which is formed from both these tags then the HEAD of that table is easily extracted. By establishing the binary matrices for the meaningful tables, the table HEADS and the table BODY is separated. For that purpose the meaningful tables are converted into binary matrices. And by extracting the HEAD the contents of the meaningful tables are displayed.

IV. SYSTEM IMPLEMENTATION

The proposed system model illustrates flow of implementation. First, in the page collection process, HTML pages are collected and their XML conversion is carried out. DOM tree is generated for the respective pages to separate every tag along with its attributes from other tags. In the second step, the transformation process is carried out which includes table recognition and table filtering operations. Third step includes the C4.5 machine learning algorithm. A threshold value of the gain ratio of decision tree algorithm decides the class of the table. i.e. meaningful tables or decorative tables. After obtaining the meaningful tables HEAD extraction operation is carried out by using binary matrices technique.

V. EVALUATION METHODOLOGY

For our experiments, we collected different web pages from internet. We carried out experiments on following modules 1) Table Filtering 2) Extracting meaningful tables using decision tree algorithm. Above mentioned modified filtering rules mainly focused on detecting decorative tables from collection of tables. We processed the total number of tables in the training data set (2697) and filtered 2154 tables.

We applied decision tree algorithm to the remaining data set. The threshold value of the gain ratio decides the classes called as meaningful tables and decorative tables. Following metrics defined below are used to measure the performance of table classifier with filtering.

\[
\text{Precision Rate(P)} = \frac{\text{CountOfCorrectTablesSystemGenerated}}{\text{TotalNumberOfTablesSystemGenerated}}
\]

\[
\text{Recall Rate(R)} = \frac{\text{CountOfCorrectTablesSystemGenerated}}{\text{Correct Generated}}
\]

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Number of tables</th>
<th>Tables Proposed</th>
<th>Correct</th>
<th>Filtered Tables</th>
<th>Correct Generated</th>
<th>Precision Rate</th>
<th>Recall Rate</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2697</td>
<td>543</td>
<td>534</td>
<td>2116</td>
<td>581</td>
<td>98.34</td>
<td>91.91</td>
<td>95.12</td>
</tr>
</tbody>
</table>

Figure 8. Performance of Table classifier with filtering

VI. OBSERVATIONS & FURTHER WORK

After applying the modified filtering rules, the ratio of separating meaningful tables from decorative tables is increases. One-dimensional tables having maximum number of images, links, whose count is similar to the count of cells present into tables, cannot gives meaningful information hence such tables are filtered out.

The further work includes processing one-dimensional tables having text data and verifying the extraction of HEAD from tables by applying different rules & using tags so as to overcome the mentioned limitations.

VII. REFERENCES