Finding Similarity between Search Results from Web Databases using Clustering Algorithm

C.Renuga*1, M.Menaka*2, V.Dineshkumar*3

* Assistant Professor, * Research Scholar
Department of Computer Science, Bharathiya Arts and Science College
Deviyakurichi, Attur, Salem, Tamilnadu, India

* Assistant Professor, Department of Computer Applications, SRMV College of Arts & Science
Periyanainkenpalayam, Coimbatore, Tamilnadu, India

Abstract— Internet has become an indispensable part of our lives. Web mining is the application of data mining process to discover knowledge from web data, including web documents, hyperlinks and web pages. The data retrieval techniques from web help to extract knowledge is used on structured data from Web data. The information collected from HTML tags and all sources, gives the minimum accuracy of result at the time of merging the collection of similar data into tables. This technique needs highest accuracy for analysis of data collection from the web. In Web information extraction, Data annotation and Data alignment are the major problems that need solutions. In Data alignment technique, the alignment is not perfect which results in some tags to be falsely detected as decorative tags, leading to incorrect grouping of the values of different attributes. The web data collection process needs to search promising predefined data alignment or data retrieval process to minimize the cleaning process of web data. This research shows the domain independent finding similarity between search results using clustering technique for rectifying the data annotation and alignment problem to great extent. The discussions of results prove that, the individual website evaluation of data and calculation among all selected websites eliminates the data alignment problem. Results promise the accuracy of data, for all the web pages used in this.

Keywords— Data Mining, Web Extraction, Annotation, Clustering, k-means, Precision, Recall.

I. INTRODUCTION

Mining, is described as extraction and integration of useful data, information and knowledge from Web page contents. The heterogeneity and the lack of structure that permeates much of the ever expanding information sources on the World Wide Web. The hypertext documents, makes automated discovery, organization, and search & indexing tools of the Internet and the World Wide Web such as Lycos, Alta Vista, WebCrawler, ALIWEB, MetaCrawler, and others provide some comfort to users, but they do not generally provide structural information nor categorize, filter, or interpret documents. In recent years these factors have prompted researchers to develop more intelligent tools as depicted in fig 1.1 for information retrieval, such as intelligent web agents, as well as to extend database and data mining techniques to provide a higher level of organization for semi-structured data available on the web.

The web content mining is differentiated from two different points of view: Information Retrieval View and Database View. Summarized the research works done for unstructured data and semi-structured data. It shows that most of the researches use bag of words, which is based on the statistics about single words in isolation, to represent unstructured text and take single word found in the training corpus as features.

II. RELATED WORK

The data alignment technique becomes emerging topic of interest. The alignment algorithm is not perfect which results in some tags to be falsely detected as decorative tags, leading to incorrect merging of the values of different attributes. From the literature, it is found that in a few cases some texts are not assigned labels by any of the annotators. One reason is that some texts are for cosmetic or navigating purposes. These texts do not represent any attributes of the real-world entity and they are not the labels of any data unit, which belong to our One-To-Nothing relationship type. It is also possible that some of these texts are indeed data units but none of current
basic annotators are applicable to them. Web information extraction and annotation has been an active research in recent years. Before extraction of the data collection from web, analysis of the web page takes important role in web mining. Data alignment is an important factor in achieving the accurate details from web data. Most existing automatic data alignment techniques are based on one or very few features. The most frequently used feature is HTML tag path finding and tree generation technique. The unused data are needed to be modified or to be removed from the original collected data. The data cleaning process have been applied for that data collection. The accuracy of the individual webpage is needed for the performance analysis. This research provides a new mining method for data collection from the web. The following drawbacks were addressed by the proposed techniques.

- False detection of decorative tags.
- Incorrect Merging of different attribution
- Automatic construction of annotation wrapper for search result records.
- Data Alignment Problem

III. METHODOLOGY

Web information extraction and annotation has been an active research area in recent years. Many systems [4] rely on human users to mark the desired information on sample pages and label the marked nodes at the same time, and then the system can induce a series of rules (wrapper) to extract the same set of information on web pages from the same source. These systems are often referred as a wrapper induction system. Because of the supervised training and learning process, these systems can usually achieve high extraction accuracy. However, they suffer from poor scalability and are not suitable for applications [5] that need to extract information from a large number of web sources.

For utilize ontology together with several heuristics to automatically extract data in multi record documents and label them. However, ontologies for different domains must be constructed manually. It’s all exploit the presentation styles and the spatial locality of semantically related items, but its learning process for annotation is domain dependent. Moreover, a seed of instances of semantic concepts in a set of HTML documents needs to be hand labeled. These methods are not fully automatic. The effort is to automatically construct wrappers are, used for data extraction only (not for annotation)[6]. There are several works, which aim at automatically assigning meaningful labels to the data units in SRRs. Basically annotated at units with the closest labels on result pages.

This method has limited applicability because many WDBs[7] do not encode data units with their labels on result pages. The existing system collects the data from web and aligns it based on content and tags with similarity. Fig 2.1 depicts the annotation wrapper for the collected web data based on the similarity.

IV. PROBLEM FORMULATION

The architecture of proposed system as depicted in 3.2 describes about the overall process of the research work. The flow of content contains the smaller works in the algorithm process. The data is extracted from the web page. The extracted data is the first phase of the research work. It checks the alignment process for storage purpose. The collected data will be stored in the local database for further calculations.

The same process is repeated for all other web pages for data retrieval. The noisy data removed after the extraction of the web data. Based on the user queries similar book data will be displayed for users using the string similarity hamming distance algorithm and grouped into same table by applying modified k-means clustering algorithm. The price threshold is added to group the cluster so as to extract the book domains with price threshold. The result contains the all related books from the stored data. The precision and recall is applied for the analyzing the content quality of the web site.

- Data unit and Text node

Each node in such a tag structure is either a tag node or a text node. A tag node corresponds to an HTML tag surrounded by “<” and “>”. Text nodes are the visible elements on the web page and data units are located in the text nodes. The four types of relationship between data unit(U) and Text node(T).

- One-to-One Relationship

In this type, each text node contains exactly one data unit, i.e., the text of this node contains the value of a single attribute. For example, each text node surrounded by the pair of tags <A> and </A> is a value of the title attribute. An automatic text node is equivalent to a data unit.

- False detection of decorative tags.
- Incorrect Merging of different attribution
- Automatic construction of annotation wrapper for search result records.
- Data Alignment Problem

III. METHODOLOGY

Web information extraction and annotation has been an active research area in recent years. Many systems [4] rely on human users to mark the desired information on sample pages and label the marked nodes at the same time, and then the system can induce a series of rules (wrapper) to extract the same set of information on web pages from the same source. These systems are often referred as a wrapper induction system. Because of the supervised training and learning process, these systems can usually achieve high extraction accuracy. However, they suffer from poor scalability and are not suitable for applications [5] that need to extract information from a large number of web sources.

For utilize ontology together with several heuristics to automatically extract data in multi record documents and label them. However, ontologies for different domains must be constructed manually. It’s all exploit the presentation styles and the spatial locality of semantically related items, but its learning process for annotation is domain dependent. Moreover, a seed of instances of semantic concepts in a set of HTML documents needs to be hand labeled. These methods are not fully automatic. The effort is to automatically construct wrappers are, used for data extraction only (not for annotation)[6]. There are several works, which aim at automatically assigning meaningful labels to the data units in SRRs. Basically annotated at units with the closest labels on result pages.

This method has limited applicability because many WDBs[7] do not encode data units with their labels on result pages. The existing system collects the data from web and aligns it based on content and tags with similarity. Fig 2.1 depicts the annotation wrapper for the collected web data based on the similarity.

IV. PROBLEM FORMULATION

The architecture of proposed system as depicted in 3.2 describes about the overall process of the research work. The flow of content contains the smaller works in the algorithm process. The data is extracted from the web page. The extracted data is the first phase of the research work. It checks the alignment process for storage purpose. The collected data will be stored in the local database for further calculations.

The same process is repeated for all other web pages for data retrieval. The noisy data removed after the extraction of the web data. Based on the user queries similar book data will be displayed for users using the string similarity hamming distance algorithm and grouped into same table by applying modified k-means clustering algorithm. The price threshold is added to group the cluster so as to extract the book domains with price threshold. The result contains the all related books from the stored data. The precision and recall is applied for the analyzing the content quality of the web site.

- Data unit and Text node

Each node in such a tag structure is either a tag node or a text node. A tag node corresponds to an HTML tag surrounded by “<” and “>”. Text nodes are the visible elements on the web page and data units are located in the text nodes. The four types of relationship between data unit(U) and Text node(T).

- One-to-One Relationship

In this type, each text node contains exactly one data unit, i.e., the text of this node contains the value of a single attribute. For example, each text node surrounded by the pair of tags <A> and </A> is a value of the title attribute. An automatic text node is equivalent to a data unit.
**One-to-Many Relationship**

It consists of three data units: Book Name, Author, and Price. Since the text of such kind of nodes can be considered as a composition of the texts of multiple data units, called composite text node.

**Many-to-One Relationship**

The value of the Author attribute is contained in multiple text nodes with each embedded inside a separate pair of (&lt;A&gt;, &lt;/A&gt;) HTML tags. This kind of tags as decorative tags because they are used mainly for changing the appearance of part of the text node. For the purpose of extraction and annotation, need to identify and removes these tags inside SRRs so that the wholeness of each split data unit can be restored.

**One-To-Nothing Relationship**

The text nodes belonging to this category are not part of any data unit inside SRRs. Further observations indicate that these text nodes are usually displayed in a certain pattern across all SRRs.

- **Data units and Text node Features**

  A data unit is a piece of text that semantically represents one concept of an entity[8]. Five common features shared by the data units belonging to the same concept across all SRRs and all of them can be automatically obtained. It is not difficult to see that all these features are applicable to text nodes, including composite text nodes involving the same set of concepts and template text nodes.

- **Data Content (DC)**

  The data units or text nodes with the same concepts often share certain keywords. This is true for two reasons. First, the data units corresponding to the search field where the user enters a search condition usually contain the search key words. Second, web designers sometimes put some leading label in front of certain data unit within the same text node to make it easier for users to understand the data.

- **Presentation Style (PS)**

  This feature describes how a data unit is display on a webpage. It consist of six style features: font style, font size, font color, font weight. Data units of the same concepts in different SRRs are usually displayed in the same style.

- **Data Type (DT)**

  Each data unit has its own semantic type although it is just a text string in the HTML code. The following basic data types are currently considered in this approach: Data, Decimal, Integer, Symbol and String. The data type of a composite text node is the concatenation of the data types of all its data units.

- **Tag Path (TP)**

  A tag path of a text node is a sequence of tags traversing from the root of the SRR to the corresponding node in the tag tree. Each node in the expression contains two parts, one is the tag name, and the other is the direction indicating whether the next node is the first child.

**V. DISCUSSIONS**

**A. Initializing the Web Database**

<table>
<thead>
<tr>
<th>S.No</th>
<th>Book Name</th>
<th>Website 1</th>
<th>Website 2</th>
<th>Website 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Author</td>
<td>Price</td>
<td>Author</td>
</tr>
<tr>
<td>1</td>
<td>Traveling the silk</td>
<td>Nil</td>
<td>Nil</td>
<td>Mar</td>
</tr>
<tr>
<td>2</td>
<td>The Soul of a New</td>
<td>Nil</td>
<td>Nil</td>
<td>Tra</td>
</tr>
<tr>
<td>3</td>
<td>Tibet: Perspectives</td>
<td>Nil</td>
<td>Nil</td>
<td>Nil</td>
</tr>
</tbody>
</table>

The extracted data stored in the local stored database is shown in Tab 5.2. The initialization process controls the flow of data at the time of data load. It is depicted as in the screen. As shown Fig 4.6 minimum time is availed to retrieve every web page from websites. It shows book details of all the websites stored in the database. The String similarity measure together with the price threshold proves that the mined results for web data bases is efficient and is user friendly.

**B. Hamming Distance for String Similarity Measure**

The accuracy measure is shown in Tab III. The search results using hamming distance for string similarity among all the books from stored data. The table shows the process of finding relevant books by splitting the book string into word for similarity match. The result shows the individual result of each website of the books stored. The total number related
books can be retrieved in that website. Before clustering, a similarity/distance measure is determined.

The measure reflects the degree of closeness or separation of the target objects. It should correspond to the characteristics that are believed to distinguish the clusters embedded in the data. In many cases, these characteristics are dependent on the data or the problem context at hand, search is no measure that is universally best for all kinds of clustering problems. Moreover, choosing an appropriate similarity measure is also crucial for cluster analysis, especially for a particular type of clustering algorithms[7].

The Hamming Distance is a number used to denote the difference between two strings. The differences between the two strings are mentioned by the numeric notation. It gives the higher performance of the comparison algorithm.

Step 1: Search Key : Perspectives
Step 2: String from Website 1

<table>
<thead>
<tr>
<th>S.No</th>
<th>Book Name</th>
<th>String Length</th>
<th>Matched Char</th>
<th>Matching Result</th>
<th>Search Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tibetan Perspectives and Prospects..</td>
<td>32</td>
<td>Perspectives</td>
<td>0 (yes)</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Indian Saris: Traditions Perspectives..</td>
<td>20</td>
<td>Perspectives</td>
<td>0(yes)</td>
<td>2</td>
</tr>
</tbody>
</table>

Step 3 : Data Cleaning Process
It removes the unwanted data in the decorative tags and it deletes the following characters of the by
The delimiters namely <, >, /, “,”, are also cleared.

Step 4: String Similarity Measure
Search String s1: Tibetan Perspectives and Prospects
Search String s2: Indian Saris: Traditions Perspectives
Using Hamming Distance algorithm, first string s1- perspectives and second string s2- perspective. So both the string is matched and it display the result in numeric. String is matched and it displays the value is 0.
Total length of the first string is 32, after data cleaning the checked from it position is 7 to 18 characters is matched the with search key. The result is yes, the number is 0. The search count is incremented by 1.

The first string match is Perspectives and second string match is idea. The characters are not matched and it displays the 1(No) is not included in the extracted results.
Step 5: Modified K-Means Clustering with price Threshold
The extracted book and its price value are taken for price threshold value.

Table – VI Mined Books with Price Threshold

<table>
<thead>
<tr>
<th>N</th>
<th>Book Name</th>
<th>Url1</th>
<th>Url2</th>
<th>Url3</th>
<th>Accuracy</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The House of Morgan: An</td>
<td>87.5</td>
<td>100</td>
<td>33.3</td>
<td>X/MAX(count)*100</td>
<td>35 ms</td>
</tr>
<tr>
<td>2</td>
<td>The Firm: The New</td>
<td>91.17</td>
<td>100</td>
<td>25.4</td>
<td>X/MAX(count)*100</td>
<td>34 ms</td>
</tr>
<tr>
<td>3</td>
<td>The Soul of a New</td>
<td>71.5</td>
<td>71.5</td>
<td>25.26</td>
<td>X/MAX(count)*100</td>
<td>34 ms</td>
</tr>
<tr>
<td>4</td>
<td>Travelling the silk</td>
<td>78.2</td>
<td>100</td>
<td>15.94</td>
<td>X/MAX(count)*100</td>
<td>34 ms</td>
</tr>
<tr>
<td>5</td>
<td>The Chief of the</td>
<td>74.1</td>
<td>100</td>
<td>25.80</td>
<td>X/MAX(count)*100</td>
<td>6/7 ms</td>
</tr>
</tbody>
</table>

Select the numeric value ‘0’ if price value is > 745 or < 945
Otherwise eliminated the book from mining.
The mined result from web database with price threshold is shown in Table.

VI. CONCLUSIONS
The data received from the web site and its stored as collected data is the result of the web mining process in the data mining. The extracted data will be used for the further research in all areas. In this research, Extracted data is aligned to the same format as to that of the data, at the time of extraction itself. Thus the inbuilt process of alignment technique is used in the proposed research. The Hamming distance string similarity measure is used, for finding nearest and relevant books from all the data. These computed similarity measure is applied for the calculation of Precision and Recall for each and individual website to compute the accuracy. The clustering process of k-means algorithm is applied for the grouping of similar domain (books name) into same cluster. The result shows that the precision and recall of this method are both above 96 percent. The results are graphically displayed by using the visualization techniques of data mining. The results prove that the extracted data are cleaned and aligned with high.

REFERENCES


