In linguistics, annotations include comments and metadata; these non-
transcriptional annotations are also non-linguistic. A collection of texts with
linguistic annotations is known as a corpus (plural corpora). The Linguistic Annota-
tion Wiki describes tools and formats for creating and managing linguistic anno-
tations.

A web annotation is an on line annotation associated with a web resource, typi-
cally a web page. With a Web annotation system, a user can add, modify or
remove information from a Web resource without modifying the resource itself.
The annotations can be thought of as a layer on top of the existing resource, and
this annotation layer is usually visible to other users who share the same annota-
tion system. In such cases, the web annotation tool is a type of social software
tool. For Web-based text annotation systems, see Text annotation.

Web annotation can be used for the following purposes:
• To rate a Web resource, such as by its usefulness, user-friendliness, suitabil-
ity for viewing by minors.
• To improve or adapt its contents by adding/removing material, something
like a wiki.
• As a collaborative tool, e.g. to discuss the contents of a certain resource.
• As a medium of artistic or social criticism, by allowing Web users to reinter-
pret, enrich or protest against institution or ideas that appear on the Web.
• To quantify transient relationships between information fragments.

2. EXISTING SYSTEM
An increasing number of databases have become web accessible through HTML
form-based search interfaces. The data units returned from the underlying data-
base are usually encoded into the result pages dynamically for human browsing.
For the encoded data units to be machine processable, which is essential for
many applications such as deep web data collection and Internet comparison
shopping, they need to be extracted out and assigned meaningful labels. In this
paper, we present an automatic annotation approach that first aligns the data units
on a result page into different groups such that the data in the same group have the
same semantic. Then, for each group we annotate it from different aspects and
aggregate the different annotations to predict a final annotation label for it. An
annotation wrapper for the search site is automatically constructed and can be
used to annotate new result pages from the same web database. Our experiments
indicate that the proposed approach is highly effective. [1]

In this paper, we studied the data annotation problem and proposed a multi anno-
tator approach to automatically constructing an annotation wrapper for annotat-
ing the search result records retrieved from any given web database. This
approach consists of six basic annotators and a probabilistic method to combine
the basic annotators. Each of these annotators exploits one type of features for
annotation and our experimental results show that each of the annotators is useful
and they together are capable of generating high quality annotation. A special fea-
ture of our method is that, when annotating the results retrieved from a web data-
base, it utilizes both the LIS of the web database and the IIS of multiple web data-
bases in the same domain. We also explained how the use of the IIS can help alle-
We see two main possible evolutions: a Web system, and its use on the Sym'Previus data warehouse. As future works, the results displayed to end users are an ordered list of tables, from the most lower/upper expectations, allowing us to reflect uncertainty in the evaluation. Finally, reliability evaluations and ordering of data tables are achieved by using belief functions, since they offer a good compromise between flexibility and comprehensiveness. It relies on the use of basic probabilistic assignments and of induced criteria. It is central in TP usage, we review methods for TP matching optimization. They are applied to decide important characteristics. This paper is a comprehensive survey of these topics, in which we outline and compare the various features of tree patterns. We also review and discuss the two main families of approaches for optimizing tree pattern matching, namely pattern tree minimization and holistic matching. We finally present actual tree pattern-based developments, to provide a global overview of this significant research topic.

We provide in this paper a comprehensive survey about XML tree patterns, which are nowadays considered crucial in XML querying and its optimization. We first compare TPs from a structural point of view, concluding that the richer a TP is with matching possibilities, the larger the subset of XQuery/XPath it encompasses, and thus the closer to user expectations it is.

Second, acknowledging that TP querying, i.e., matching a TP against a data tree, is central in TP usage, we review methods for TP matching optimization. They belong to two main families: TP minimization and holistic matching. We trust we provide a good overview of these approaches' evolution, and highlight the best algorithms in each family as of today. Moreover, we want to emphasize that TP minimization and holistic matching are complementary and should both be used to wholly optimize TP matching.

3. PROPOSED ALGORITHM

Annotating the web search results are very useful these days for the users for many reasons such as comparison of available products, on-line shopping, articles etc. This work is proposing a mechanism for fast searching of web data for articles using annotations applied automatically for future usage for the various websites visited by the users of the system. The complete algorithm shall be implemented using following steps:

Step 1: Load the website in the system

Step 2: Retrieve the articles from it using various links available on the pages of the website or load the separate pages if available directly.

Step 3: Look for heading tags, bold/strong tags, and frequency of the words in the articles to decide the annotation for the specific articles

Step 4: Various Intermediate steps for data cleaning, filtering, stemming and stopping are applied to decide important characteristics.

Step 5: Frequencies of the various characteristics based categories of the documents are evaluated for calculating various metric values such as True Positive, True Negative, False Positive, False Negative, Precision and Recall for measuring the accuracy.

Step 6: Time requirements of the step by step processes have been recorded for measuring the performance.

Step 7: Results are tabulated and graphs have been drawn for the various readings and comparison is done with the existing works.

1) Algorithms Applied for annotating the articles:
   - Data Unit and Text Node Feature Extraction
   - Data Content
   - Presentation Style
   - Data Type
   - Tag Path
   - Adjacency

2) Data Alignment Techniques

3) Annotators
   - Frequency Based Annotators
4. RESULTS & DISCUSSION

Table 1: Readings Taken from Simulation of the Proposed Work

<table>
<thead>
<tr>
<th>SNO</th>
<th>CATEGORY</th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
<th>PRECISION</th>
<th>RECALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Course</td>
<td>1251</td>
<td>152</td>
<td>7596</td>
<td>256</td>
<td>89.2</td>
<td>83</td>
</tr>
<tr>
<td>2</td>
<td>Project</td>
<td>1633</td>
<td>452</td>
<td>7214</td>
<td>0</td>
<td>78.3</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>Department</td>
<td>230</td>
<td>25</td>
<td>8822</td>
<td>50</td>
<td>90.0</td>
<td>82.1</td>
</tr>
<tr>
<td>4</td>
<td>Faculty</td>
<td>1125</td>
<td>50</td>
<td>8797</td>
<td>558</td>
<td>95.7</td>
<td>66.8</td>
</tr>
<tr>
<td>5</td>
<td>Other</td>
<td>923</td>
<td>50</td>
<td>8797</td>
<td>302</td>
<td>94.8</td>
<td>75.3</td>
</tr>
<tr>
<td>6</td>
<td>Staff</td>
<td>742</td>
<td>43</td>
<td>8804</td>
<td>113</td>
<td>94.5</td>
<td>86.8</td>
</tr>
<tr>
<td>7</td>
<td>Student</td>
<td>600</td>
<td>50</td>
<td>8797</td>
<td>46</td>
<td>92.3</td>
<td>92.9</td>
</tr>
</tbody>
</table>

Fig 1: Recall Value calculated from the proposed work

Fig 2: Precision Value calculated from the proposed work

Fig 3: Time Taken during calculation of the proposed work

Comparison with the Existing Work

Table 2: Comparison on the basis of various characteristics

<table>
<thead>
<tr>
<th>CHARACTERISTIC</th>
<th>EXISTING WORK</th>
<th>PROPOSED WORK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work Type</td>
<td>Comparison of the Various Annotation Techniques and Alignment Techniques</td>
<td>Semantic Annotations with Frequency Annotation Methodology</td>
</tr>
<tr>
<td>Dataset Applied</td>
<td>4 Different Test Data</td>
<td>Web Knowledge Base Dataset containing data of 7 Categories for Various Universities</td>
</tr>
<tr>
<td>Annotation Term Used</td>
<td>Single</td>
<td>All (fetched from the dataset)</td>
</tr>
<tr>
<td>Annotation Term Decision</td>
<td>Manual</td>
<td>Automatic</td>
</tr>
<tr>
<td>Alignment Technique Required</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Performance Evaluated</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Accuracy Measures (Precision &amp; Recall Applied)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Size of Data Set</td>
<td>40 Documents</td>
<td>245 Documents</td>
</tr>
</tbody>
</table>

Table 3: Comparison Chart for Precision and Recall of the Existing & Proposed Algorithm

<table>
<thead>
<tr>
<th>SNO</th>
<th>Algorithm</th>
<th>Precision (Average)</th>
<th>Recall (Average)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Existing</td>
<td>83.0</td>
<td>59.0</td>
</tr>
<tr>
<td>2</td>
<td>Proposed</td>
<td>90.7</td>
<td>83.8</td>
</tr>
</tbody>
</table>

Fig 4: Comparison between Existing & Proposed Work

The proposed work is being implemented using C# for websites of the blogs where annotation is required to a great extent using standard machines.

REFERENCES

2. Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng, Clement Yu, “Annotating Structured Data of the Deep Web”, This work is supported in part by the following NSF grants: IIS-0414981, IIS-0414939 and CNS-0454298.


