hard to maintain. It gets difficult for the users to access number of web sites individually to get the needed information. H. He et al.[10] proposed WISE-Integrator tool which performs integration of Web Interfaces of Search Engines automatically. It is used for identifying the similar attributes from distinct search interfaces for integration. WISE-Integrator is capable of automatically grouping elements into logical attributes and deriving a rich set of meta-information for each attribute.

J.Zhu et al.[11] proposed Hierarchical Conditional Random Field approach. Current approach makes use of decoupled strategies. The data record detection and attributes labelling is done in two separate phases. It gets ineffective the idea of extracting data records and attributes separately. It proposes a probabilistic model to perform both processes simultaneously. HCRF can integrate all useful features by learning by their importance, and it can also integrate hierarchical interaction. Its limitations are cost and template dependency.

J. Wang et al.[12] proposed DeLa, a method which is very similar to proposed annotation work. DeLa's alignment method is based on HTML tags, on the other hand proposed work uses other features such as text content, adjacency information, data type, proposed annotation method deals with relationships between text nodes and data units, DeLa utilizes different search interfaces of WDBs for annotation.

3. Proposed Work

This paper considers how to assign labels to the data values present in the SRRs automatically. From a collection of search result record which have been extracted from a result page returned from a web database.

A. Objectives

- Perform data extraction.
- Perform data alignment

B. System flow

Components of system flow includes SRR's, data extraction, data alignment, combing tag value similarity, SRR results. The system flow is shown below in fig.2,

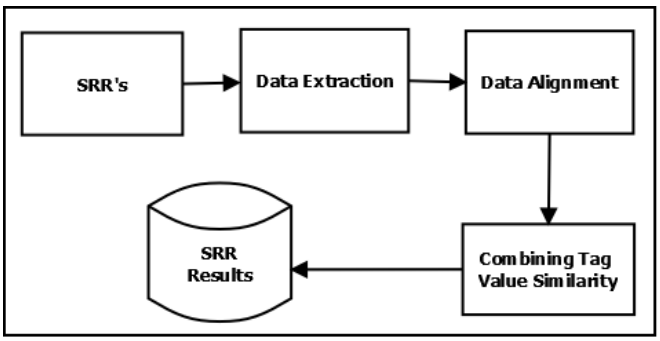

Figure 2: System Flow

4. Data Extraction

For the annotation purpose, the search result records have to be extracted from the returned result pages. So as the irrelevant information such as links, information about the hosting site and advertisements are to be discarded from the result page. It is very time consuming and tedious for extracting the data records from the result page by manually written programs and it is not practical as the search engine changes the result page display over period of time. So as to extract the records from the results page a wrapper generator which bases on Visual information and Tag structure is used. The extraction of data using ViNT is based on features such as visual content of the web page and also the HTML tag structure of the page in HTML format.

Each search result record is saved in a tag tree structure with one root and every node in the tag tree corresponds to an HTML tag in the original page.Fig.3 shows the tag tree structure of a html page, using this structure, it becomes easy to find each node in the HTML page. The information like physical position , its coordinates and area size of each node can also be obtained using ViNTs.

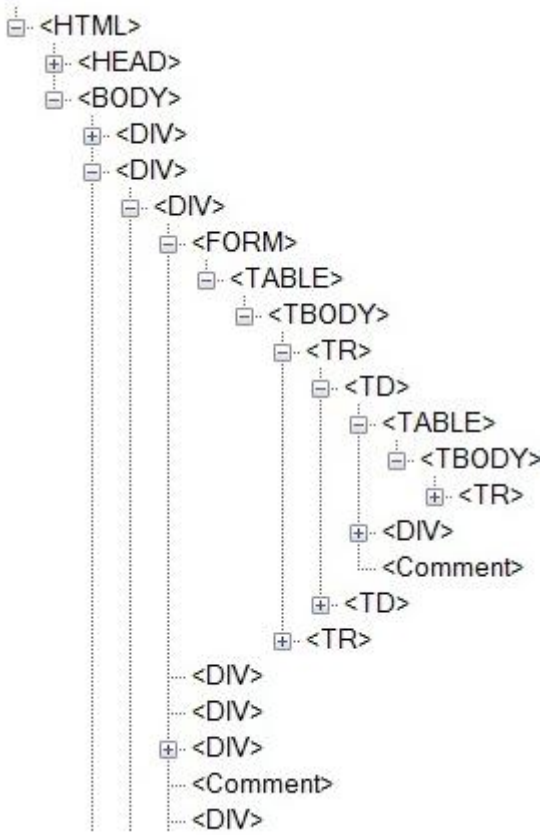

Figure 3: Tag tree structure of a webpage

5. Data Alignment

Once the extraction of search result record from the web page is done, the data units are generally not aligned. The aim of alignment of data is to place the data units from search result records into a group which has same semantic. Alignment of records makes annotation of data a lot easier. It is based on the idea that the data units in different search results of the same semantic mostly have the determined presentation and layout.

Data Alignment concentrates on five features for similarity such as Data Unit, Data Content, Presentation Style, Data Type, Tag Path Similarity which is described in [8].So as to

increase the efficiency of data grouping and alignment, a cluster based shifting technique is used.
Alignment is carried out based on same features, a group of data units which are having similar features are put in one group by aligning it. If a group contains data units of one concept and if there is no other data unit of another concept then the group is known as well aligned group. The aim of alignment is to put the data units in the table so every alignment group is well aligned.

The goal of data alignment is to put the data units of the same concept into one group so that they can be annotated comprehensively.

Alignment Steps:

a) Merge Text Nodes
It detects and removes decorative tags from every SRR, which permits the text nodes identical to the same attribute to be merged into a single one.

b) Align Text Nodes
After merging, it aligns text nodes into different groups. So that same group has the same concepts.

c) Split (Composite) Text Node
In this step the composite text nodes are splitted into separate data unit.

d) Align Data Units
This is the last step for alignment, in which every composite groups are separated in different multiple aligned groups, which contains data units of same concept.

Alignment Algorithm:

1) Read Source HTML file which contains the records.
2) Process each record in html nodes in the source html file one by one.
3) For every "Node" in the "Root", check if the element contains data or empty node.
4) If element contains "data node", then we are going to consider them as fully qualified records, which can be used in accessing for search process.
5) If element doesn't contain "data node", which might be missing in construction of document which is of no use we are going to eliminate them for further processing.

Clustering algorithm :

1) Read all fully available records from the annotation stage.
2) For each record evaluate all the "child node", and if child nodes contain full data then those records will be taken high distance records.
3) Non-available "child nodes" will be pushed in to the last part of SRR generation.4.
4) When user performs "search" SRR's will be processed according to fully available data.

6. Data alignment, labelling and wrapper generation:

Automatic annotation is based on alignment approach which aligns the data units by using different types of relationship in between data units and text nodes. A cluster-based shifting algorithm is used in alignment process. After the successful alignment label the data units and automatically construct an annotation wrapper for the search site.

7. Experimental Results

The fig.4 shows the performance graph of the system, the experiments performed on the various html datasets and results that are emerged ,the results shows that the implemented technique outperforms the previous work done.

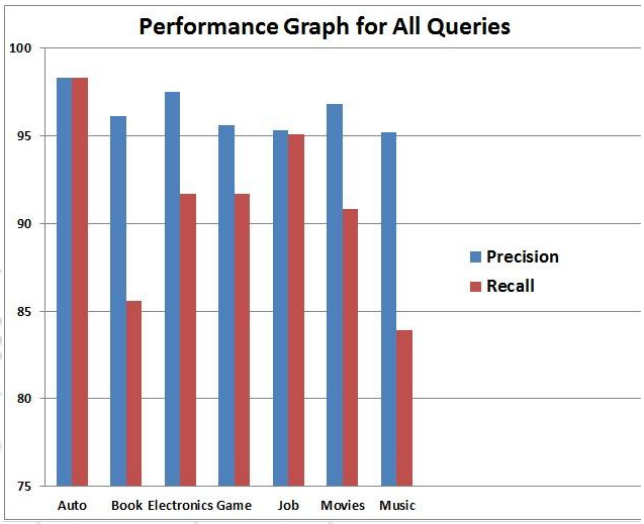

Figure 4: Performance Results for HTML datasets

8. Conclusion

In this paper, the data annotation problem is mentioned and Implemented a multi-annotator approach to annotate the SRR'S, an automatic annotation wrapper is used to search result records retrieved from web database. A new data extraction method is implemented so as to extract search result records automatically from a result page. It uses two steps for this task, first it includes identifying and segmenting the search result records. Existing methods are improved by allowing the SRR'S in a data region to be non-adjoining. In second step it aligns the data values among the SRR's. A unique alignment method is implemented in which the alignment is performed pairwise and nested structure processing. Experimental result shows that CTVS is more accurate and performs better.

References

[1] Arasu and H. Garcia-Molina, "Extracting Structured Data from Web Pages," Proc. SIGMOD Int'l Conf. Management of Data, 2003.
[2] V. Crescenzi, G. Mecca, and P. Merialdo, "RoadRUNNER: Towards Automatic Data Extraction from Large Web Sites," Proc.Very Large Data Bases (VLDB) Conf., 2001.
[3] J. Wang, J. Wen, F. Lochovsky, and W. Ma, "Instance-BasedSchema Matching for Web Databases by Domain-Specific Query Probing," Proc. Very Large Databases (VLDB) Conf., 2004.
[4] L. Arlotta, V. Crescenzi, G. Mecca, and P. Merialdo, "Automatic Annotation of Data Extracted from Large

Web Sites," Proc. Sixth Int' Workshop the Web and Databases (WebDB), 2003.

[5] S. Mukherjee, I.V. Ramakrishnan, and A. Singh, "BootstrappingSemantic Annotation for Content-Rich HTML Documents," Proc.IEEE Int'l Conf. Data Eng. (ICDE), 2005.

[6] W. Su, J. Wang, and F.H. Lochovsky, "ODE: Ontology-Assisted Data Extraction," ACM Trans. Database Systems, vol. 34, no. 2,article 12, June 2009.

[7] D. Embley, D. Campbell, Y. Jiang, S. Liddle, D. Lonsdale, Y. Ng, and R. Smith, "Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages," Data and Knowledge Eng., vol. 31,no. 3, pp. 227-251, 1999.

[8] W. Liu, X. Meng, and W. Meng, "ViDE: A Vision-Based Approach for Deep Web Data Extraction," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 3, pp. 447-460, Mar. 2010.

[9] H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu, "Fully Automatic Wrapper Generation for Search Engines," Proc. Int'l Conf. World Wide Web (WWW), 2005.

[10] H. He, W. Meng, C. Yu, and Z. Wu, "Automatic Integration of Web Search Interfaces with WISE-Integrator," VLDB J., vol. 13,no. 3, pp. 256-273, Sept. 2004.

[11] J. Zhu, Z. Nie, J. Wen, B. Zhang, and W.-Y. Ma, "Simultaneous Record Detection and Attribute Labeling in Web Data Extraction," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2006.

[12] J. Wang and F.H. Lochovsky, "Data Extraction and Label Assignment for Web Databases," Proc. 12th Int'l Conf. WorldWideWeb,2003.

[13] N. Krushmerick, D. Weld, and R. Doorenbos, "Wrapper Induction for Information Extraction," Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI), 1997.

[14] L. Liu, C. Pu, and W. Han, "XWRAP: An XML-Enabled Wrapper Construction System for Web Information Sources," Proc. IEEE 16th Int'l Conf. Data Eng. (ICDE), 2001.

[15] W. Meng, C. Yu, and K. Liu, "Building Efficient and Effective Metasearch Engines," ACM Computing Surveys, vol. 34, no. 1, pp. 48-89, 2002.

[16] Z. Wu et al., "Towards Automatic Incorporation of Search Engines into a Large-Scale Metasearch Engine,"