INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN: 2278 – 0211 (Online)

Ontology Based Data Unit Similarity With Combining Tag And Value For Data Extraction And Alignment

K. Jeyalakshmi
Assistant Professor, PG And Research Department Of Computer Science
Hindusthan College Of Arts And Science, Tamilnadu, India
Anitha J
Research Scholar, PG And Research Department Of Computer Science
Hindusthan College Of Arts And Science, Tamilnadu, India

Abstract:
Web database extraction is used to retrieve relevant information from the query result page. By combining tag and value one can extracts data from query result pages by first identifying and segmenting the query result records (QRRs) in the query result pages and then aligning the segmented QRRs into a table. But combining tag and value similarity measure doesn't handle non-contiguous QRR. To overcome this problem a novel method is proposed to display the most distinct query records from user's query result pages. In this method, First distinct tags are extracted from the result records to build the tag vector table, and then the similarity between each record is found using several similarity methods. Finally the values of similar records are combined and aligned using ontology based alignment.

Key words: *Ontology based CTVS, Web Database Extraction, Jaccard Similarity Measure, QRR Extraction, Distinct Tag and Value Extraction using ontology*

1.Introduction

Information Extraction refers to the automatic extraction of structured information such as entities, relationships between entities and attributes describing entities from unstructured sources. This enables much richer forms of queries on the abundant unstructured sources than possible with keyword searches alone. Due to the presence of auxiliary information such as a comment, recommendation or advertisement QRRs are not contiguous. Almost all other data extraction methods assume that the QRRs are presented contiguously in only one data region in a page. However this assumption may not be true for many web databases where auxiliary information separates the QRRs. In some web pages, redundant information's are displayed. To avoid redundant information and to optimize the result page with only relevant information ontology based data unit similarity method is proposed. Ontology based CTVS processes other similarity function that measure the presentation style based similarity and data content after the results are returned for each user query. In CTVS the data value and presentation style similarity information effectively returns most similar data results. Because it shares similar tag structures, a flat structure with several columns having the same tag structure might be mistakenly identified as a nested structure. But In Ontology based CTVS while combing the result records the similar records are identified and eliminated to display by using similarity measures results.

2.Methodologies

Ontology data is interpreted as a set of objects and relationships between them. It applies an attribute identifier function (e.g., pattern matching, keyword-based search) to locate an attribute's occurrences in the data source. The proposed system introduce the additional attribute similarity function while measure the data similarity from the return result pages in the records that related to user query. The approach automatically extracts the query result records (QRRs) from HTML pages (to a user query) dynamically generated by a deep web site. Only when the data are extracted and stored in a database they can are easily compared and aggregated using traditional database querying techniques. Consequently, an accurate data extraction method is vital for these applications to operate correctly. The goal is therefore to acquire sufficient domain knowledge from the query interfaces and query result pages in a domain and to use the acquired knowledge to extract the data instances from the query result pages of the domain.
Component 1: Ontology construction for a given domain (fully automatic).
Component 2: Data Extraction using the ontology (fully automatic).

3.Ontology Based Contruction

In ontology based construction system first the result (QRR) in the same domain automatically extracted by measuring the similarity also. By measuring the data unit similarity with presentation, data unit, data based similarity using the pages returned in the web sites. Uses both the query interfaces and the query result pages of web sites from the same domain to automatically construct domain ontology.

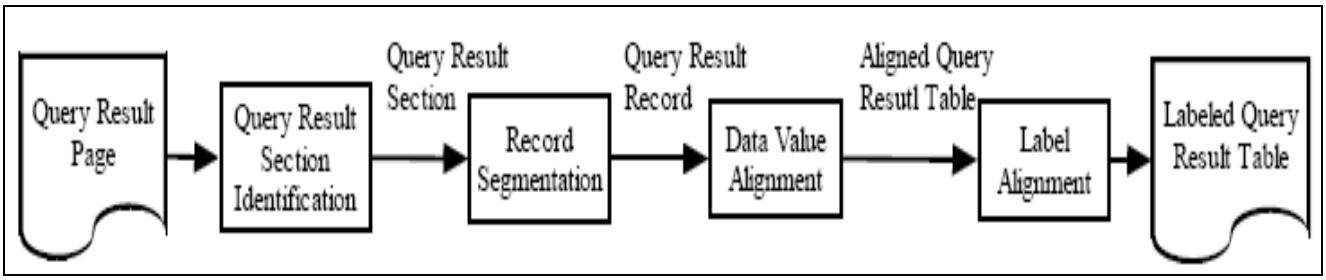

Figure 1: The Steps Of Ontology

The steps of the ontology

Identifies query result section in a query result page using the ontology.

Segments the query result section into query result records (QRRs), and Aligns and labels the data values in the query result records into a table so that the data values for the same attribute in each record are put into the same column in the table.

Query result section identification Is done by constructing a tag tree from the query result pages and finding a sub tree whose raw data strings have a large correlation with the ontology.

Record segmentation technique same as primary wrapper (according to the starting and ending positions of the tandem repeats and the visual gap between the segmented records).

Data value alignment and label assignment both tasks can be performed simultaneously by assigning an attribute name of the ontology to each data value of a QRR. A maximum entropy model is used to do attribute name assignment for data values.

4.Similarity Measures

1. Data Type Similarity (SimD)

It is determined by the common sequence of the component data types between two data units. The longest common sequence (LCS) cannot be longer than the number of component data types in these two data units. Thus, let t_1 and t_2 be the sequences of the data types of d_1 and d_2, respectively, and TLen(t) represent the number of component types of data type t, the data type similarity between data units d_1 and d_2 is

$$SimD(d_1, d_2) = \frac{LCS(t_1, t_2)}{Max(Tlen(t_1).Tlen(t_2)}$$

2. Data Content Similarity (SimC)

It is the Cosine similarity between the term frequency vectors of d_1 and d_2:

$$SimC(d_1, d_2) = \frac{V_{d_1}, V_{d_2}}{\left||V_{d_1}|\right| * \left||V_{d_1}|\right|}$$

Where V_d is the frequency vector of the terms inside data unit d, $||V_d||$ is the length of V_d, and the numerator is the inner product of two vectors.

3. Presentation Style Similarity (SimP)

It is the average of the style feature scores (FS) over all six presentation style features (F) between d_1 and d_2

$$SimP(d_1, d_2) = \sum_{i=1}^{6} FS_i/6$$

Where FS_i is the score of the i^{th} style feature and it is defined by $FS_i = 1$ if $F_d^1 = F_d^2$ and $FS_i = 0$ otherwise, and F_d^I is the i^{th} style feature of data unit d

4. Jaccard Similarity

The Jaccard similarity is defined as the quotient between the intersection and the union of the pairwise compared variables among two objects.

$$d^{jac}(i,j) = \frac{f11}{f01 - f10 + f11}$$

In the equation d^{JAD} is the Jaccard distance between the objects i and j. For two data records with n binary variables y the variable index k ranges from 0 to n-1.

implementation

1. CONSTRUCTION OF TAG PATH:
Given a query result page, it constructs a tag tree for the page rooted in the <HTML> tag. Each node represents a tag in the HTML page and its children are tags enclosed inside it. Each internal node n of the tag tree has a tag string t_{s_n}, which includes the tags of n and all tags of n's descendants, and a tag path t_{p_n}, which includes the tags from the root to n.

2. DATA REGION IDENTIFICATION:
It identifies all possible data regions which usually contain dynamically generated data in top down fashion starting from the root node. It is recursively applied to the children of n_i only if it does not have any similar siblings. Multiple data regions may be identified in this step.

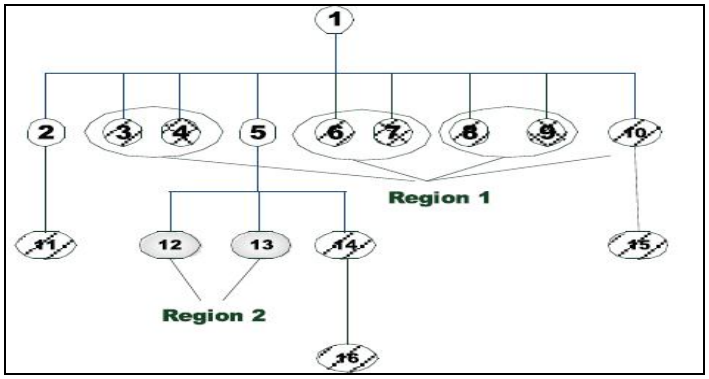

Figure 2: Region 1, Region 2= No Child Node To Same Parent

Data region identification algorithm discovers data regions in a top-down manner. Starting from the root of the query result page tag tree, the data region identification algorithm is applied to a node n and recursively to its children n_i, i = 1...m as follows: Compute the similarity sim_{ij} of each pair of nodes n_i and n_j;i ;j = 1...m and i $\neq$ j, using the node similarity calculation method. The data region identification algorithm is recursively applied to the children of n_i only if it does not have any similar siblings. For example in figure algorithm has to be applied to the children of node 5 since it does not have any similar sibling node. The recognized similar nodes with the same parent form a data region.

3. SEGMENTATION:
In Segmenting the QR Page, The following two heuristics are used for the tandem repeat selection:
1. If there is auxiliary information, which corresponds to nodes between record instances, within a data region, the tandem repeat that stops at the auxiliary information is the correct tandem repeat since auxiliary information usually is not inserted into the middle of a record.

Tandem repeats: Two similar nodes with repetitive subparts and considered as dissimilar.
Region 1 represented as ABABABA if we use characters A to represent an element of the similar node set {3, 6, 8, 10} and B to represent an element of the similar node set {4, 7, 9}. In this case, there are two tandem repeats AB and BA. Similarly Region 2 in Fig. 3 can be represented as CC, which contains only one tandem repeat, C.

4. SIMILARITY MEASURE:
In similarity measures are
1) Data type similarity
2) Data content similarity
3) Presentation style similarity
4) Jaccard Similarity is measured.

These similarities are used to find the quotient between the intersection and the union of the pair wise compared variables among two objects on ontology basis. The result of measured value is stored in the vector table. Using vector table a threshold value is set to find and extract the distinct result records.

5. DATA REGION MERGE:

Given the segmented data records, the Data Region Merge module merges the data regions containing similar records. The similarity between any two records from two data regions is measured by the similarity of their tag strings with threshold used to judge whether two records are similar.

Given columns cp in a holistic alignment and a similarity threshold S_{nest} as input, the procedure nested decides, using the similarities of the data values in c_p, whether cp contains a repetitive tag pattern that is formed by a nested structure. We assume that two columns are generated by the same attribute if there is a large data value similarity between these two columns. Given a column c_1, which contains m data values, we define the intra column similarity sim_{intra} to be the average data value similarity within each column in c_1

$$sim_{intra} = \frac{2\sum_{j=1}^{m-1}\sum_{i=j+1}^{m} s_{ij}}{m(m-1)}$$

s_{ij} is the data value similarity between the i^{th} and j^{th} data values of c1. For c_p, its intra column similarity is the average of the intra column similarity of all columns in c_p

For two columns c_1 and c_2, which have m and n data values, respectively, the inter column similarity sim_{inter} is defined to be the average data value similarity of every pair of data values in c_1 and c_2

$$sim_{inter} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{m} s_{ij}}{mn}$$

5.Performance Analysis

RECALL CALCULATION:

RECALL values are the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage.

The graph shows the result for record level calculation. The experiment result shows that proposed system gives the best recall result while comparing to CTVS. As the number of records increases also the proposed system got the best recall result than base paper Nested structure.

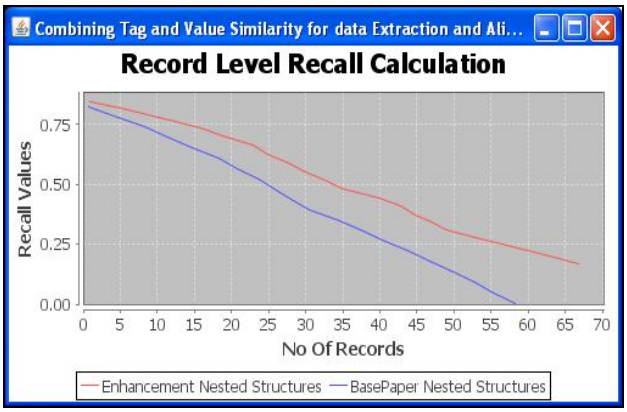

Figure 3: Recall Calculation

PRECISION CALCULATION:

PRECISION is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage.

RECORD LEVEL PRECISION CALCULATION:

If the number of records increases also our precision rate is high when compared to CTVS.

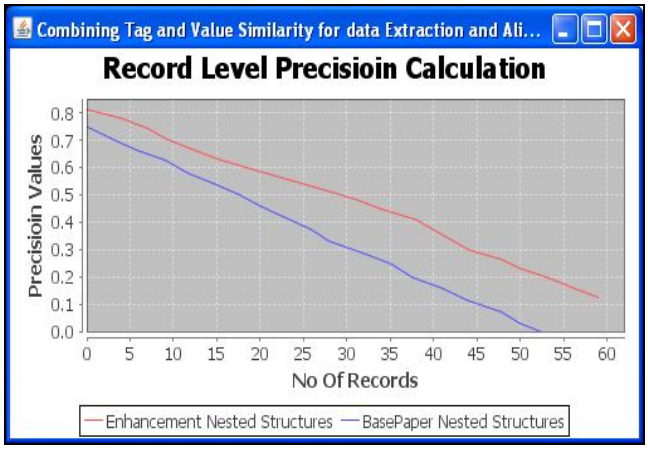

Figure 4: Record Level Precision Calculation

PAGE LEVEL PRECISION CALCULATION:
If the number of pages increases also our precision rate is high when compared to CTVS.

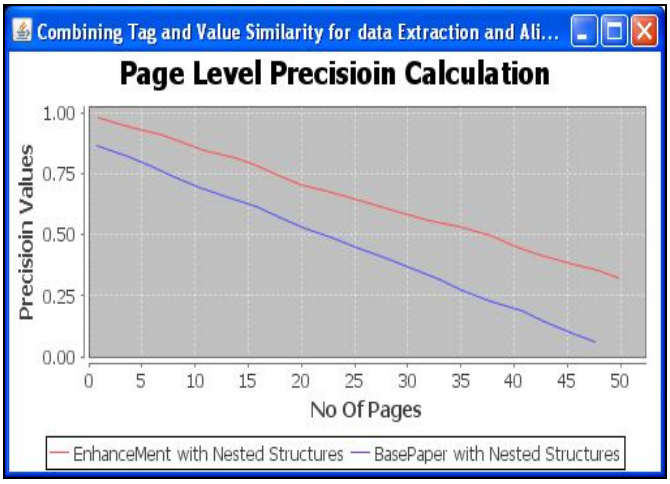

Figure 5: Page Level Precision Calculation

6.Conclusion

Ontology based data unit similarity with combining tag and value for data extraction and alignment, system Determine whether there is another query result section that contains QRRs and return the query result section if it exists. It removes the section that has been incorrectly identified as the query result section from the query result page and identifies a new section in the remaining query result page by measuring the similarity such as data, content type and presentation style similarity. In proposed system a new efficient ontology based mining is developed to handle any nested structure in the QRRs after the holistic alignment and the issue of ambiguity arise in learning-based approaches to schema inference and data extraction from Web documents has been resolved. Finally it avoids the redundant query result records. The future work may be extended on how efficiently it can extract the data from web pages based on the notion of structural-semantic entropy.

7.Acknowledgment

I express my sincere thanks to all people who have contributed a lot for the successful implementation of this work. I take this opportunity to express my deep sense of gratitude to our Management Trustees, Hindusthan Educational and Charitable Trust, Coimbatore for providing abundant facilities to carry out my research work successfully on the campus.

I convey my sincere thanks to Dr. N. Balusamy, M.Cop, MBA, Ph.D, D.Ed, Principal, Hindusthan College of Arts and Science, Coimbatore, for his constant support and guidance. I express my deep sense of gratitude to Mr. R. Rangaraj, M.Sc, M.Phil, M.Sc(Psychology), (Ph.D), Head, PG and Research Department of Computer Science, Hindusthan College of Arts and Science, Coimbatore, for his keen interest, support and suggestion.

I really deem this as a special privilege to convey my prodigious and everlasting thanks to my supervisor Mrs. K. Jeyalakshmi, MCA, M.Phil, (Ph.D), Assistant Professor, PG and Research Department of Computer Science, Hindusthan College of Arts and Science, Coimbatore, for her valuable guidance and personal interest in my research work.

8.References

1. A.V. Goldberg and R.E. Tarjan, "A New Approach to the Maximum Flow Problem," Proc. 18th Ann. ACM Symp. Theory of Computing, pp. 136-146, 1986.
2. R. Baeza-Yates, "Algorithms for String Matching: A Survey," ACM SIGIR Forum, vol. 23, nos. 3/4, pp. 34-58, 1989.
3. D.W. Embley, D.M. Campbell, Y.S. Jiang, S.W. Liddle, D.W. Lonsdale, Y.-K. Ng, and R.D. Smith, "Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages," Data and Knowledge Eng., vol. 31, no. 3, pp. 227-251, 1999.
4. Mobasher, B., Cooley, R. and Srivastava, J. "Automatic Personalization based on web usage Mining" Communications of the ACM, Vol. 43, No.8, pp. 142–151, 2000.
5. N. Kushmerick, "Wrapper Induction: Efficiency and Expressiveness," Artificial Intelligence, vol. 118, nos. 1/2, pp. 15-68, 2000.
6. L. Liu, C. Pu, and W. Han, "XWRAP: An XML-enabled Wrapper Construction System for Web Information Sources," Proc. 16th Int'l Conf. Data Eng., pp. 611-621, 2000.
7. R. Baumgartner, S. Flesca, and G. Gottlob, "Visual Web Information Extraction with Lixto," Proc. 27th Int'l Conf. Very Large Data Bases, pp. 119-128, 2001.
8. M.K. Bergman, "The Deep Web: Surfacing Hidden Value," White Paper, BrightPlanet Corporation, http://www.brightplanet. com/ resources/details/ deepweb. html, 2001.
9. P. Bonizzoni and G.D. Vedova, "The Complexity of Multiple Sequence Alignment with SP-Score that Is a Metric," Theoretical Computer Science, vol. 259, nos. 1/2, pp. 63-79, 2001.
10. D. Buttler, L. Liu, and C. Pu, "A Fully Automated Object Extraction System for the World Wide Web," Proc. 21st Int'l Conf. Distributed Computing Systems, pp. 361-370, 2001.
11. C.H. Chang and S.C. Lui, "IEPAD: Information Extraction Based on Pattern Discovery," Proc. 10th World Wide Web Conf., pp. 681- 688, 2001.
12. W. Cohen and L. Jensen, "A Structured Wrapper Induction System for Extracting Information from Semi-Structured Documents," Proc. IJCAI Workshop Adaptive Text Extraction and Mining, 2001.
13. V. Crescenzi, G. Mecca, and P. Merialdo, "Roadrunner: Towards Automatic Data Extraction from Large Web Sites," Proc. 27th Int'l Conf. Very Large Data Bases, pp. 109-118, 2001.
14. Muslea, S. Minton, and C. Knoblock, "Hierarchical Wrapper Induction for Semistructured Information Sources," Autonomous Agents and Multi-Agent Systems, vol. 4, nos. 1/2, pp. 93-114, 2001.
15. W. Cohen, M. Hurst, and L. Jensen, "A Flexible Learning System for Wrapping Tables and Lists in HTML Documents," Proc. 11th World Wide Web Conf., pp. 232-241, 2002.
16. J. Wang and F. Lochovsky, "Data-Rich Section Extraction from HTML Pages," Proc. Third Int'l Conf. Web Information System Eng., 2002.
17. Eirinaki, M., Vazirgiannis, M. "Web Mining for Web Personalization", ACM Transactions on Internet Technology, Vol.3, No.1, February 2003.
18. Arasu and H. Garcia-Molina, "Extracting Structured Data from Web Pages," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 337-348, 2003.
19. Liu, R. Grossman, and Y. Zhai, "Mining Data Records in Web Pages," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 601-606, 2003.
20. K.C.-C. Chang, B. He, C. Li, M. Patel, and Z. Zhang, "Structured Databases on the Web: Observations and Implications," SIGMOD Record, vol. 33, no. 3, pp. 61-70, 2004.
21. L. Chen, H.M. Jamil, and N. Wang, "Automatic Composite Wrapper Generation for Semi-Structured Biological Data Based on Table Structure Identification," SIGMOD Record, vol. 33, no. 2, pp. 58-64, 2004.