International Journal of Emerging Trend in Engineering and Basic Sciences (IJEEBS)
*ISSN (Online)*2349-6967
Volume 2 , Issue 1(Jan-Feb 2015), PP281-286

PAGE RANKING BASED ON COMBINING TAG AND VALUE SIMILARITY

Pawan Kote[1], Saurabh Munde[2], Vikas Jagadale[3], Chaitali Thakur[4]Prof.Supriya Bhosale[5]
[1 2 3 4 5](*Dept. of Information Technology, Dr.D.Y.Patil College of Engg, Pune, India*)

Abstract: -An automatic web record extraction extracts a set of objects from heterogeneous web pages based on similarity measure among objects in an automated fashion. Automatically extracting the data from these query result pages is very important for many applications, such as integration of data, which need to work with multiple databases. This classifies a region in the web page according to similar data object which emerge frequently in it. The transformation of unstructured data into structured data that can be stored and analyzed in a central local database is involved. The existing system develops a data extraction and alignment method known asPGCTVS, which can identify the query result records (QRRs) by extracting the data from query result page and then segment them. We also design a new record alignment algorithm that aligns the attributes in a new record, first pair wise and then holistically. Applications of the extraction of data records also includes the concept of page ranking in order to speed up the search engine, clustering of the similar data regions is performed for efficient identification of web pages and similar Query result pages are applicable also in data integration and comparison shopping.

Keywords: - Automatic wrapper generation, data record alignment, false positive rate detection, informationintegration, page ranking.

I. INTRODUCTION

The World Wide Web is large collection of information on growing demand. It is complicated to query in unstructured data. The paper focuses on the issues of automatically extracting data that are encoded in the query result pages generated by databases. It does not require any human input like manually generated rules or training sets.We present a novel data extraction and alignment method called page ranking based on combining tag and value similarity that combines both tag and value similarity. Many web applications, such as Integrity of data and shop comparison, need the data from multiple web databases. There are three stages to extract objects from a Web page. It includes record extraction, attribute alignment and attribute labeling. In general, a query result page (QRP) contains not only the actual data, but also other additional information such as navigable panels, ads, reviews, data about site to be presented, and so on. We employ the following two-step method,called Page Ranking Based on Combining Tag and Value Similarity (PGCTVS), to extract the QRRs from a query result page p.

1. Record extraction processidentifies the QRRs
2. Record alignment aligns the data values of the QRPs in p.

For example, Fig.shows a query result page fragment containing two QRRs for Transcends product whichspecifies both the queries by their <size >, <price>.

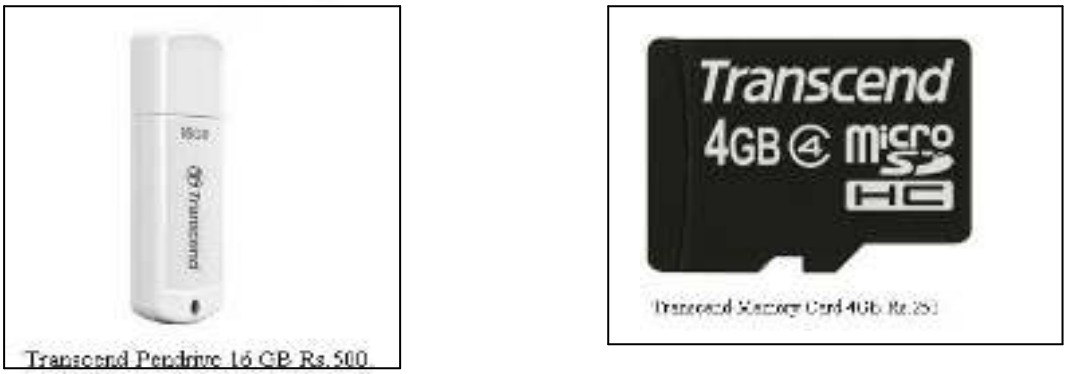

Fig. 1 An example query result page for query—Brand: Transcend.

II. LITERATURE SURVEY

2.1. Web Data Extraction and Alignment Tools: A Survey

It is published in International Journal of Scientific Engineering and Technology Volume No.2, Issue No.6,pp : 573-578 written by Shridevi A. Swami, PujashreeVidap .
This paper deals with the study of various automatic web data extraction and data alignment techniques.
Advantage:-Wrapper Induction viz. As user itself labels the items of interest, no extra data are extracted.

Disadvantage:-Manual labelling of data is time-consuming. It is not scalable to a large number of web databases.

2.2. ODE: Ontology-assisted Data Extraction

W. Su, J. Wang, and F.H. LochovskyACM Trans. Database Systems, vol. 34, no. 2, article 12, p. 35, 2009.
ODE which automatically extracts the query result records from the HTML pages. It first constructs an ontology for a domain according to query interfaces or query result pages within a domain. However, there are some attributes whose labels never appear in any query interface or query result page. QRR alignment is performed by a novel three-step data alignment method that combines tag and value similarity.

Advantage:-ODE can effectively process query result pages containing zero or few QRRs. Fully automatic data extraction — All data extraction steps can be performed automatically (i.e., no manual labeling or other human interaction is required).[1]

Disadvantage:-To label attributes it is necessary that the label appeared in query interfaces or query result pages within a domain.

III. QRR ALIGNMENT

1.Pairwise QRR alignment aligns the data values in a pair of QRRs to provide the evidence for how the data values should be aligned among all QRRs
2. Holistic alignment aligns the data values in all the QRRs.
3.Nested structure processing identifies the nested structures that exist in the QRRs.

3.1 Apriori Algorithm Processing

Aprioriis an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database.[6]

Apriori algorithm is shown in below Fig. 2

Join Step: Ckis generated by joining Lk-1with itself
•Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset
•Pseudo-code:

Ck: Candidate itemset of size k
Lk: frequent itemset of size k
L1= {frequent items};
for(*k*= 1; *Lk*!=∅; *k*++) **do begin**
Ck+1= candidates generated from *Lk*;
for each transaction *t*in database do
increment the count of all candidates in *Ck+1*that are contained in *t*
Lk+1= candidates in *Ck+1*with min_support
end
return∪*kLk*;

Fig 2.Apriori Algorithm

Example 1.A large supermarket tracks sales data by Stock-keeping unit (SKU) for each item, and thus is able to know what items are typically purchased together. Apriori is a moderately efficient way to build a list offrequent purchased item pairs from this data. Let the database of transactions consist of the sets {1,2,3,4},{1,2,3,4,5}, {2,3,4}, {2,3,5}, {1,2,4}, {1,3,4}, {2,3,4,5}, {1,3,4,5}, {3,4,5}, {1,2,3,5}. Each number corresponds to a product such as "butter" or "water". The first step of Apriori is to count up the frequencies, called the supports, of each member item separately:

Item	Support
1	6
2	7
3	9
4	8
5	6

Item	Support
{1,3}	5
{1,4}	5
{1,5}	3
{2,3}	6
{2,4}	5
{2,5}	4
{3,4}	7
{3,5}	6
{4,5}	4

We can define a minimum support level to qualify as "frequent," which depends on the context.For this case, let min support = 4. Therefore, all are frequent. The next step is to generate a list of all 2-pairsof the frequent items. Had any of the above items not been frequent, they wouldn't have been included as a possible member of possible 2-item pairs. In this way, Aprioriprunes the tree of all possible sets. In next step we again select only these items (now 2-pairs are items) which are frequent (the pairs written in bold text).
We generate the list of all 3-triples of the frequent items (by connecting frequent pair with frequent single item).

TABLE 1

Item	Support
{1,3,4}	4
{2,3,4}	4
{2,3,5}	4
{3,4,5}	4

The algorithm will end here because the pair {2,3,4,5} generated at the next step does not have the desired support.
We will now apply the same algorithm on the same set of data considering that the min support is
5. We get the following results:

Step 1:

Item	Support
1	6
2	7
3	9
4	8
5	6

Step 2:

Item	Support
{1,2}	4
{1,3}	5
{1,4}	5
{1,5}	3
{2,3}	6
{2,4}	5
{2,5}	4
{3,4}	7
{3,5}	6
{4,5}	4

The algorithm ends here because none of the 3-triples generated at Step 3 have de desired support.
3.2 Data Value Similarity Calculation

Given two data values f1 and f2 from different QRRs, werequire their similarity, s12, to be a real value in [0, 1]. Thedata value similarity is calculated according to the data type tree shown in Fig. 4. Each child node is a subset of its parent node.

For example, the "string" type includes several children data types, which are common on the web such as "datetime," "float," and "price." The maximum depth of the data type tree is 4. In the following, we will refer to a nonstring data type as a specific data type.

Given two data values f1 and f2, we first judge their data types and then fit them as deeply as possible into the nodes n1 and n2 of the data type tree. For example, given a string "784," we will put it in node "integer."

The similarity s12 between two data values f1 and f2 with data type nodes n1 and n2 is defined as

$$s_{12} = \begin{cases} 0.5 & n_1 = p(n_2)\ \&\ (n_1) \neq OR \\ & p(n_1)\ \&n_2 \neq \text{String} \\ 1 & n_1 = n_2 \neq \text{String} \\ \text{cosine similarity} & n_1 = n_2 \neq \text{String} \\ 0 & \text{otherwise,} \end{cases}$$

where p(ni) refers to the parent node of ni in the data type tree. The similarity between data values f1 and f2 is set to 0.5, if they belong to different specific data types that have a common parent. 1, if they belong to the same specific data type.string cosine similarity of f1 and f2, if both f1 and f2
belong to the string data type. 0 otherwise, which occurs when one of f1 and f2 belongs to the string data type and the other one belongs to a specific data type, or f1 and f2 belong to different specific data types without any direct parent.

IV. DESIGN CONSIDERATION

4.1 System Flow

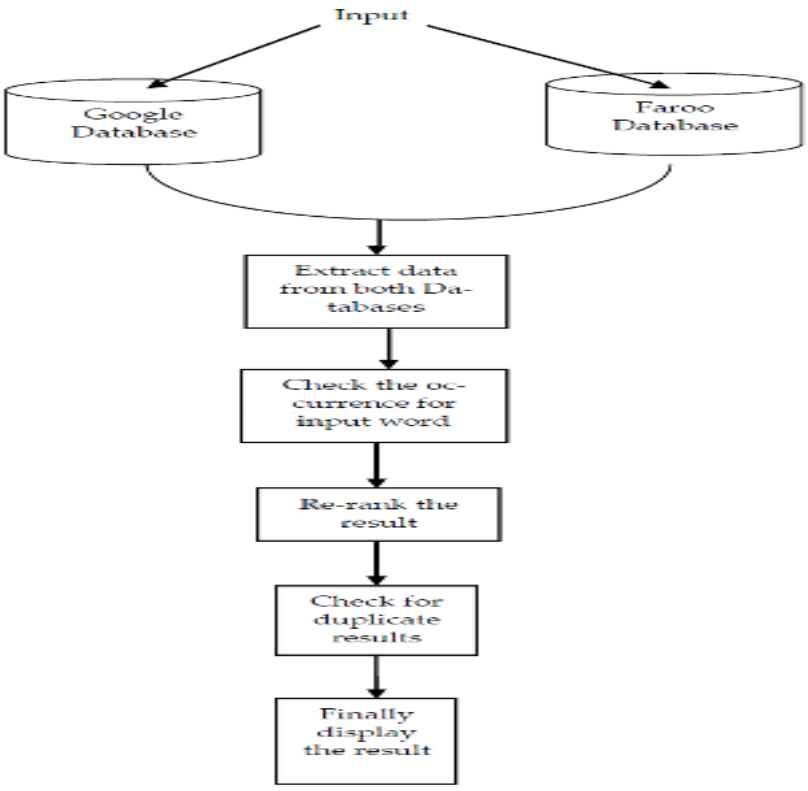

Fig 3. System Architecture

Entered query is searched in Google and Faroo database, the entered query is extracted from both databases. Then the query is checked for occurrence in Server. If the query entered is frequent search query then it is reranked. It will also verify for duplicate values and make it a structured query.[4]

TABLE 2: Data Extraction Method

Method	Nested Structure Processing	Single Result Page	Non Contiguous Data Region
PGCTVS	Yes	Yes	Yes
DeLa	Yes	Yes	No
ViNTs	No	No	No

4.2 Evaluation Setup

We have executed different tests to note the performance of updated Page Ranking Based on Combining Tag and Value Similarityalgorithm. We now present the experimental results for updated Page Ranking Based on Combining Tag and Value Similarity over five data sets and compare Updated PGCTVS with ViNTs, and DeLa. We have chosen ViNTs and DeLa to compare with PGCTVS because both have been shown to perform very accurate data extraction and implementations of both are available to us. An Updated PGCTVS is implemented in .Net. When running on a Pentium 2.5 GHz CPU with 1 GB memory, the running time required to process a page is 0.087 second son average computed over a random selection of 100 pages. We have also used page ranking method in order to bring out the expected data records so that the time consumption is less. Correct extraction of data records has been tested and the accurate results have been achieved. The performance of the data extraction methods is compared in three different ways. The other two evaluations focus on specific properties of the query result pages. Noncontiguous QRR evaluation differentiates the performance for query result pages in which the QRRs are contiguous and noncontiguous.

V. CONCLUSION

The Existing Data Extraction Method (PGCTVS) allows the Query Result Records in a data region to be non-contiguous as well as aligns the data values among the QRR. It has been shown to be an working data extraction method it does not figure out the case where multiple data values from more than one attribute are clustered inside one leaf node of the tag tree and data value of a single attribute spans multiple leaf nodes. Here the proposed structural-semantic entropy is calculated for each node in a DOM tree. It focus on identifying data-rich regions and find the lowest common parent nodes of the sibling subtrees forming the records in the DOM tree representation of a web page with the help of a set of keywords. The future work may be extended to extract the data from web pages based on the design issues such as amount of memory usage, computational workload, process etc. The algorithm used requires that the entropy should be calculated for every non-leaf node of a DOM tree. One of the possible approach is to find rules to terminate the calculation before the entropies of all nodes are calculated in a bottom-up way. On the other hand increase speed of the data extraction for the pages during the process of crawling a website.

VI. FUTURE WORK

Page ranking method has been introduced based on giving the weight to the particular webpage for speeding up the extraction of web page process. Clustering algorithm is to stipulate the clustering of related web pages so as to reduce the collision that has been taken place in the previous PGCTVS method. We improved our algorithm with these existing techniques by allowing the QRRs in a data region to be non-contiguous. A novel alignment method is introduced in which the alignment is performed in three consecutive steps: pair wise allocation, holistic allocation, and nested structure operation. Experiments on five sets show that PGCTVS is generally more accurate than current state-of-the-art methods. Although PGCTVS has been shown to be an accurate data extraction method, it still has some boundaries. First, it requires at least two QRRs in the query result page. Future enhancements of page ranking method and clustering methods and the time consumption in retrieving the data has been reduced accordingly Website linking structure has been identified and implemented in order to find the linkage between the web pages. Future enhancement are the page ranking or algorithm has been constructed also if a campaign of exchanging links to increase Page Rank is to be implemented, it is crucial that the importance of factors such as link text is understood. Page Rank declares that a link from not a occasionally

visited and rarely updated web page should not have equal weighting to a link from a popular web page. The purpose of the Page Rank algorithm is to include a score, between zero to ten, to every web site.

REFERENCES

[1] W. Su, J. Wang, and F.H. Lochovsky, "ODE: Ontology-AssistedData Extraction," ACM Trans. Database Systems*, vol. 34, no. 2,article 12, p. 35, 2009*

[2] Weifeng Su, Jiyang Wang, Frederick H. Lochovsky, "Combining Tag and Value Similarity for Data Extraction and Alignment" IEEE Transactions on Knowledge and Data Engineering*, vol.24, No.7, July 2012.*

[3] Y. Zhai and B. Liu, "Structured Data Extraction from the WebBased on Partial Tree Alignment," IEEE Trans. Knowledge and DataEng.*, vol. 18, no. 12, pp. 1614-1628, Dec. 2006.*

[4] Duhan, N. "Page Ranking Algorithms: A Survey" Advance Computing Conference,*2009. IACC 2009.IEEE International.*

[5] L. Chen, H.M. Jamil, and N. Wang, "Automatic Composite Wrapper Generation for Semi-Structured Biological Data Based on Table Structure Identification,"SIGMODRecord,*vol. 33, no. 2, pp. 58-64, 2004*.

Books:

[6] Jiawei Han, *Data Mining Concepts and Techniques*(Simon Fraser University)

[7] Thomas Connolly, *Database Systems* (Addison Wesley)

[8] Korth, *Database System Concepts* (McGrawHill)