International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 4, April - 2013

Incorporating Optional Labeling And Dynamic Tagging With Combining Tag And Value Similarity

Vishnupriya. G. Warrier
PG Scholar,
Dept. of Computer Science and Engineering
Anna University of Technology
Regional Centre, Coimbatore

D. Palanikkumar
Assistant Professor,
Dept. of Computer Science and Engineering
Anna University of Technology
Regional Centre, Coimbatore

Abstract

Query result pages are generated from the web databases based on the query given by the user. From these query result pages data are extracted automatically. The data extraction is based on the combined tag and value similarity technique. The Query Result Record (QRRs) in the query result pages are identified and segmented. The segmented QRRs are then aligned in to a table. The non contiguous QRRs are also considered which is induced by the auxiliary information. We propose new techniques called Optional Labeling and Dynamic Tag structuring which improves the efficiency in data extraction. Initially all the tags are stored temporarily in a database, from where relevant tags are extracted. The tag structuring is handled dynamically so that more accurate extraction is made possible.

Index terms: CTVS, Data extraction, data record alignment, information integration, wrapper generation

1. Introduction

Online databases are supported by the deep web.In deep web[16],the pages are dynamically generated in response to the user's query where as in a surface web unique URLs are needed.Relevant data encoded in HTML pages are generated by web database as a result of the user's query. For better performance,accurate data extraction is necessary. For many web applications automatic data extraction is a mandatory fact. The data extraction should be in a structured manner[1].

This paper focuses on the problem of considering optional attributes also for data extraction. For this process label assignment technique is referred[12]. Another proposal is the introduction of dynamic tagging. The goal of extraction is the removal of any irrelevant information from the query result page, considering the optional attribute conditionally, extracting the Query result records[4] and aligning those QRRs into a table[17].

We introduce a new technique called Optional Combined Tag and Value Similarity(OCTVS) for the extraction of QRRs from a query result page.

1) *Record extraction* identifies the QRRs in a query result page which involve the following substeps:data region identification,buffering,semantic extraction and the segmentation step.
2) *Record Alignment* where the data values for the same attribute are aligned and put in to the same column of the table

Comparing with the existing CTVS technique , OCTVS improves the data extraction accuracy in 2 ways:

1) *Optional labeling* is the technique by which the problem of elimination of optional attribute that appears as the start node in a data region ,as an auxiliary information is eliminated.This is incorporated in the record extraction step.
2) *Dynamic tagging* is the other improvement.The existing system uses static tagging which results in less accurate results.The dynamic tagging uses the semantic data extraction concept.In the static tagging only the attributes and values recorded in prior can be used.

IJERT

Wrappers are used to extract search result records from the result pages that are dynamically generated by search engines [8]. Wrapper building consists of various sub processes such as identifying the candidate search result records(SRRs), finding the tag paths of records, wrapper format hypothesis, initial building, refining, selection of wrappers and wrapper integration.

Each path node *pn* consists of two components, the tag name and the direction. If the next node following the *pn* on the path is the next sibling of *pn* then it is indicated by 'S', and if the first child of *pn*, then it is indicated by 'C'.

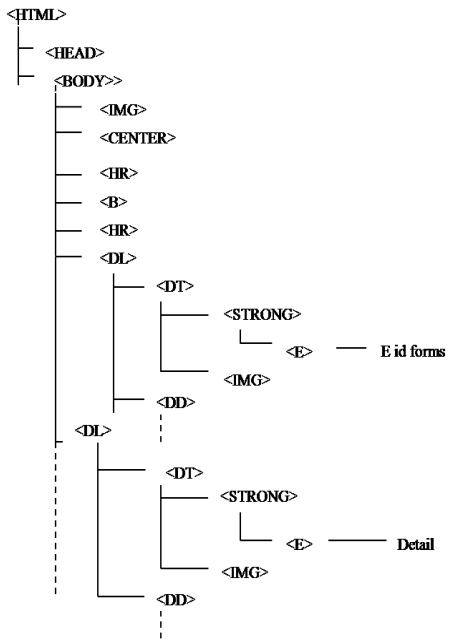

Figure 1. Tag record path

As an example the tag path of the first <IMG> in Figure 1 is "<HTML>C<HEAD>S<BODY>C"; and the tag path of the first link(E id forms);<E> is part of the link is "<HTML>C<HEAD>S<BODY>C<IMG>S<CENTER >S<HR>S<B>S<HR>S<DL>C<DT>C<STRONG>C".

Identification of data path is an inevitable step of our process. The information can be stored and retrieved accurately if and only if the tag paths are identified exactly. Once if the paths are identified, result records are generated based on the type of attributes. The attributes in a result record should be aligned. For this we use a three step process [17].

2. Related Work

Majority of us are addicted to browsing.It is the responsibility of the developer to provide the users with the accurate information without missing even the finest information.For this the hidden values[5] need to be surfaced.We refer [16] for the interested reader.

Structured objects[14] called data records present the necessary information about their host pages.Those records have to be mined, from where useful information can be extracted.Mining data records[2] mines contiguous as well as noncontiguous data records.

Information can be extracted automatically[13] by a two fold technique.Initial step includes identifying the rank potential repeating patterns with respect to the user's visual perception of the web page. Next is the alignment of data and text content.

The database values can be extracted automatically [6]without any learning examples or any human input. A set of template generated pages are taken as input and deduces the unknown template to generate the output with the values encoded in the pages.

The search engine dynamically generates a result page containing result records when a query is submitted to it. This includes irrelevant information as well. The wrapper [9, 10] can be automatically produced [8] to extract search result records. This is made possible with the utilization of both the visual content features and the HTML tag structures

Earlier wrapper induction methods require human assistance to build a wrapper. Data extraction methods for automatic extraction of records from the query result pages have been proposed recently.

Inductive learning based extraction rules are used in wrapper induction. Given a set of training pages or a list of data records, the user labels or marks the items of interest and the system learns the wrapper rules to extract records from new pages.

In wrapper induction, no extraneous data are extracted, but the manual labeling of data is tedious. So it is not recommended for large number of web data bases. Existing wrapper has poor dynamic adaptation of query result pages. To overcome the problems, some unsupervised learning methods such as Omni,IEPAD,DeLa have been proposed. These methods rely entirely on tag structures.

DeLa models the structured data as string instances encoded in HTML tags. The main problem of this method is the generation of multiple patterns and it is hard to decide the correct one.

HTML tags alone can be used to derive accurate wrappers for the following reasons:

- HTML tags are used in unexpected ways.
- Little semantic information can be conveyed using HTML tags.
- Data containing embedded tags confuse the wrapper generators making them even less reliable.

To overcome these problems methods like ViPER and ViNTs are introduced.

IJERT

Both the visual data similarity and HTML tag structures are used by ViPER. ViPER suffers in case of nested structured data, while CTVS handles the problem precisely and effectively.ViNTs learns a wrapper from a set of training pages using both visual and tag features.

3. QRR Extraction

The frame work for QRR extraction is given in figure 2. Initially a tag tree is constructed for the page rooted in the <HTML> tag[11, 15]. Each node is the representation of a tag in the HTML page and the tags enclosed in it are the children.All possible data regions are identified in the data region identification module which contains dynamically generated data in a top down fashion starting from the root node.Buffering is the storage of the identified regions in a temporary file which is further filtered during the process of considering the auxiliary information.Segmentation module segments the filtered data regions that are identified based on the tag patterns.Dynamic tagging is supported by semantic based data extraction.

Each QRR contains the information such as title of the web page,data,various attributes and their values as shown in the table1.

TABLE 1
Identified QRRs

QRR1	Data1	Title	Attr1	Val1	Attr2	Val2
QRR2	Data2	Title	Attr1	Val1	Attr2	Val2

Table 1. Identified QRRs

A. Data Region Identification

Similar data records of the same parent node are grouped as a data region.It deals with noncontiguous data records as well.Here we propose a new method which considers the auxiliary information leading to accurate data extraction.For this we need a temporary file which buffers the attributes and their values.These are further filtered conditionally to identify the exact data regions.

B. Record Segmentation

In a tag tree[18] the tandem repeats within a data region is initially found out.If only one repeat is found out it corresponds to a record.In case of multiple repeats any one has to be selected.

Heuristics for record segmentation

1) Within a data region if any auxiliary information is encountered,the tandem repeat that stops is the correct one since the auxiliary information cannot be inserted in the middle of a record.
2) If the above two heuristics are filed to be used,the tandem repeat that starts the data region is selected.

C. Data Region Merge

In a query result page there may be several data regions that are identified. Actual data records may be the integration of several data regions. Thus a thorough check is needed inorder to determine whether any of the data regions needs to be merged before the actual identification of the QRRs in a query result page.

The similarity of two data regions are considered based on the similarity of the segmented records. For this tag strings of the records of the two data regions are compared. If the average similarity is greater than or equal to 0.6 which is the threshold value[17], then the two regions can be merged.

D. Query Outcome Section Identification

There may be multiple data regions still residing in a query result page even after the merging of data regions. We are supposed to find at most one data region containing the actual QRRs.

The section identification of the query result is the same as in CTVS[17]. An area weight is assigned for each data region 'd' which is calculated as d's area divided by the largest area of all the identified regions. The query outcome section is generally located at the center of the query result page.The system architecture is shown in the figure 2.

4. System Architecture

The system architecture is as shown in figure 2.The query details are the input for the OCTVS process. The query details are the input for the OCTVS process which generates the QRR results. The QRRs are stored in a database so that it can be retrieved as per user's query. For the similarity check the tabulated information recorded in prior are checked. The extracted information are initially stored in a temporary data base . The temporary database is checked recursively to make sure that the information retrieved includes the necessary auxiliary information.

The conditional check for the elimination of auxiliary information is based on the OCTVS process . The process repeats till the exact information is retrieved as per the user's query.Finally the QRR results are stored to a database which can be further used for the merging of data regions.

IJERT

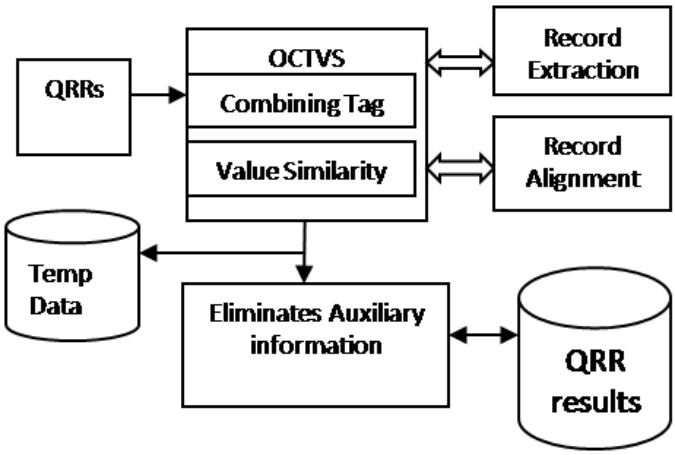

Figure 2. QRR framework for extraction

5. OCTVS Algorithm

Input : Query Result Record,R
Output: Extracted Data,E

1. Input Query
2. From the available links find the keywords
3. Store the information to a database
4. Perform structure analysis
5. Extract tags from the link
6. Store them to a temporary file
7. Match the attributes
   Identify the data regions
8. Segment the records
   Temp→ Containing optional data
   QRR→ Actual records
9. Merge QRRs
10. If the result not found then go for semantic extraction
11. Repeat step 5
12. Final Result section is identified

Algorithm 1. OCTVS Algorithm

A. Algorithm Explanation

When the user gives some query the search engine searches available links. Among those links the keywords are identified and that information is stored in a database for further extraction. Then a structure analysis is performed which includes dynamic tagging. Next step is the extraction of tags from the link. The extracted tags are stored in to a temporary file. The attributes are matched to identify the data regions. Now the records are segmented by storing the actual records in to a QRR table. Similar QRRs are merged inorder to provide the exact outcome. If the expected outcome is not found go for further tag extraction[7]. Finally the result section is identified.

6. QRR Alignment

The data values that belong to the same attribute generally show similarity in data values and may include similar strings.Data value similarity [17]is calculated between every pair of values.The pairwise alignment determines whether the paired data values belong to the same attribute on the basis of calculated data value similarity.Similarity of record path is a constraint.The alignment of data values between two QRRs must be unique.There should not be any cross alignment aswell.After the pairwise alignment all data values of the same attribute are put in to the same table column globally by means of holistic alignment.This is similar to finding connected components in an undirected graph.Vertices from the same record are not included in the same component.If any vertices breach this constraint,a breach path is to be established.However the connected components are not allowed to intersect each other.Finally a nested processing is needed to handle the attributes having multiple values.

7. Implementation and Results

The implementation is done using the ASP.NET Frame work which simplifies the application development.An application site is developed and for that site the data extraction based on the user's query is regulated using the OCTVS algorithm.Whenever a new site is to be registered the information is stored to the data base.From the stored information users are provided with the desired sites after the filtering process.Only the registered sites can be compared and retrieved.For the extraction the key word matching is the most recommended strategy.An application site developed so as to show the administrator activity of tag extraction.

Whenever a site is loaded the HTML page is analyzed and the tags are extracted and stored in a temporary data base.This process is done by administrator.Final table of records is generated after filtering the temporary data base based on the OCTVS algorithm.As the user enters a query in the application site the QRRs are analyzed and the appropriate site information is delivered with a high percentage of accuracy.

The accuracy levels can be compared by comparing the the two sites,one implemented using the OCTVS algorithm and the other implemented without the OCTVS extraction.As the user give the query the related site is delivered to the user.Both the results are compared. The one with OCTVS delivers the more accurate information than the one without OCTVS extraction.

The data extraction summarisation is given in the table 2 as a comparison with the existing system.

IJERT

TABLE 2
Data Extraction Summarization

Method	Non-contiguous data regions	Considering auxiliary information conditionally	Dynamic tag structuring
OCTVS	√	√	√
CTVS	√	×	×

8. Conclusion and Future Work

We have proposed a novel method for data extraction called OCTVS.This includes identifying the data regions which are buffered to a temporary file.The filtered data regions are segmented and finally merged before it is delivered for the selection of QRR.The alignment of records is the final step.

Although OCTVS has the advantage of providing more accurate extraction using optional labeling and dynamic tagging it has some drawbacks.It requires atleast two QRRs in the query result page.It will select only one among the data regions and discard others if a query result page has more than one data region having result records and the records in various data regions defer each other.

References

[1] Arasu and H. Garcia-Molina, "Extracting Structured Data from Web Pages," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 337-348, 2003.

[2] B. Liu, R. Grossman, and Y. Zhai, "Mining Data Records in WebPages," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 601-606, 2003.

[3] B. Liu and Y. Zhai, "NET - A System for Extracting Web Data from Flat and Nested Data Records," Proc. Sixth Int'l Conf. Web Information Systems Eng., pp. 487-495, 2005.

[4] C.H. Chang and S.C. Lui, "IEPAD: Information Extraction Based on Pattern Discovery," Proc. 10th World Wide Web Conf., pp. 681-688, 2001.

[5] C. Tao and D.W. Embley, "Automatic Hidden-Web Table Interpretation by Sibling Page Comparison," Proc. 26th Int'l Conf.

[6] 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.

[7] H. Snoussi, L. Magnin, and J.-Y. Nie, "Heterogeneous Web Data Extraction Using Ontologies," Proc. Fifth Int'l Conf. Agent-Oriented Information Systems, pp. 99-110, 2001.Conceptual Modeling, pp. 566-581, 2007.

[8] H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu, "Fully Automatic Wrapper Generation for Search Engines," Proc. 14thWorld Wide Web Conf., pp. 66-75, 2005.

[9] I. Muslea, S. Minton, and C. Knoblock, "Hierarchical WrapperInduction for Semistructured Information Sources," AutonomousAgents and Multi-Agent Systems, vol. 4, nos. 1/2, pp. 93-114, 2001.

[10] I. Muslea, S. Minton, and C. Knoblock, "A Hierarchical Approach to Wrapper Induction," Proc. Third Ann. Conf. Autonomous Agents,pp. 190-197, 1999.

[11] J. Wang and F. Lochovsky, "Data-Rich Section Extraction from HTML Pages," Proc. Third Int'l Conf. Web Information System Eng.,2002.

[12] J. Wang and F.H. Lochovsky, "Data Extraction and LabelAssignment for Web Databases," Proc. 12th World Wide Web Conf.,pp. 187-196, 2003

[13] K. Simon and G. Lausen, "ViPER: Augmenting AutomaticInformation Extraction with Visual Perceptions," Proc. 14th ACMInt'l Conf. Information and Knowledge Management, pp. 381-388,2005.

[14] K.C.-C. Chang, B. He, C. Li, M. Patel, and Z. Zhang, "StructuredDatabases on the Web: Observations and Implications," SIGMODRecord, vol. 33, no. 3, pp. 61-70, 2004.

[15] L. Liu, C. Pu, and W. Han, "XWRAP: An XML-enabled WrapperConstruction System for Web Information Sources," Proc. 16th Int'lConf. Data Eng., pp. 611-621, 2000.

[16] M.K. Bergman, "The Deep Web: Surfacing Hidden Value," WhitePaper, BrightPlanet Corporation, http://www.brightplanet.com/resources/details/deepweb.html, 2001.

[17] Weifeng Su, Jiying Wang, Frederick H.Lochovsky and Yi Liu, "Combining Tag and Value Similarity for Data Extraction and Alignment,"IEEE Transactions on Knowledge and Data Engineering,vol. 24, no. 7, pp. 1186-1199, 2012.

[18] 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.

IJERT